The American Economic Review

PAPERS AND PROCEEDINGS

OF THE

Ninety-Sixth Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

San Francisco, California, December 28–30, 1983

Program Arranged by Charles L. Schultze

Papers and Proceedings Edited by John G. Riley and Wilma St. John

MAY 1984

THE AMERICAN ECONOMIC ASSOCIATION

p14420

•Published at George Banta Co., Inc. Menasha, Wisconsin. The publication number is ISSN 0002-8282.

•*THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the Directory, and Supplements, is published by the American Economic Association and is sent to all members five times a year: March; May; June; September; December.

Dues for 1984, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature*, are as follows:

$33.00 for regular members with rank of assistant professor or lower, or with annual income of $15,840, or less;

$39.60 for regular members with rank of associate professor, or with annual income of $15,840 to $26,400;

$46.20 for regular members with rank of full professor, or with annual income above $26,400;

$16.50 for junior members (registered students). Certification must be submitted yearly.

Institutional subscriptions (libraries, institutions, or firms) are $100.00 a year. Individual subscriptions are $65.00 a year. Only subscriptions to both publications will be accepted. Single copies of either journal may be purchased from the Secretary's office.

In countries other than the United States, add $9.20 to cover extra postage.

•Correspondence relating to the Directory, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

•Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

•Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.

Founded in 1885

Officers

President

CHARLES L. SCHULTZE
The Brookings Institution

President-Elect

CHARLES P. KINDLEBERGER
Massachusetts Institute of Technology

Vice Presidents

ZVI GRILICHES
Harvard University

ROBERT L. HEILBRONER
New School for Social Research

Secretary

C. ELTON HINSHAW
Vanderbilt University

Treasurer

RENDIGS FELS
Vanderbilt University

Managing Editor of The American Economic Review

ROBERT W. CLOWER
University of California-Los Angeles

Managing Editor of The Journal of Economic Literature

MOSES ABRAMOVITZ
Stanford University

Executive Committee

Elected Members of the Executive Committee

ANN F. FRIEDLAENDER
Massachusetts Institute of Technology

JOSEPH E. STIGLITZ
Princeton University

WILLIAM D. NORDHAUS
Yale University

A. MICHAEL SPENCE
Harvard University

VICTOR R. FUCHS
Stanford University

JANET L. NORWOOD
Bureau of Labor Statistics

EX OFFICIO Members

GARDNER ACKLEY
The University of Michigan

W. ARTHUR LEWIS
Princeton University

THE AMERICAN ECONOMIC REVIEW

VOL. 74 NO. 2 MAY 1984

PAPERS AND PROCEEDINGS

OF THE

Ninety-Sixth Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

San Francisco, California

December 28–30, 1983

Program Arranged by Charles L. Schultze

Papers and Proceedings Edited by John G. Riley and Wilma St. John

CONTENTS

PAPERS

Richard T. Ely Lecture

Economic Education

Economic Effects of the Rise in Energy Prices: What Have We Learned in Ten Years?

The Determinants of Interest Rates: Old Controversies Reopened

Microeconomics of Innovation and Productivity Growth

Psychological and Sociological Foundations of Economic Behavior

Keynes, Schumpeter, Marx Centenary Session

Economic Theory and Management Education

The Political Economy of Budget Deficits

International Differences in Wage and Price Behavior

Is the United States Underinvesting in Public Capital?

What Survives of the Rational Expectations Revolution?

Information and Macroeconomics

Has the Response of Inflation to Macro Policy Changed?

Limits to Growth: Resources, Productivity, and Social Aspects

Tax Policy: A Further Look at Supply-Side Effects

Gender Roles in Economic Life

The Appropriate Response to Trade Barriers and "Unfair" Trade Practices in Other Countries

Race, Crime, and Employment

A Decade of Greater Variability in Exchange Rates: An Assessment

New Evidence on the Demand for Safety and the Behavioral Response to Safety Regulation

The Theory of Contracts and Agency: Implications for the Behavior of Wages

Recent Developments in the Theory of Industrial Organization: Some Macroeconomic Implications

Antitrust Policy and Strategic Behavior

Monetarism: Lessons from the Post-1979 Experiment

Reflections on the Current State of Macroeconomic Theory

PROCEEDINGS

THE purpose of the American Economic Association, according to its charter, is the encouragement of economic research, the issue of publications on economic subjects, and the encouragement of perfect freedom of economic discussion. The Association as such takes no partisan attitude, nor does it commit its members to any position on practical economic questions. It is the organ of no party, sect, or institution. People of all shades of economic opinion are found among its members, and widely different issues are given a hearing in its annual meetings and through its publications. The Association, therefore, assumes no responsibility for the opinions expressed by those who participate in its meetings. Moreover, the papers presented are the personal opinions of the authors and do not commit the organizations or institutions with which they are associated.

Editors' Introduction

This volume contains the *Papers and Proceedings* of the ninety-sixth annual meetings of the American Economic Association.

The *Proceedings* record the business activities of the Association in 1983: the annual membership meeting; the March and December meetings of the Executive Committee; reports of the Association's officers and committees.

The *Papers* constitute the greater part of the volume. They comprise seventy-six contributions that fill roughly the same number of pages as two regular issues of the *American Economic Review*. About a year in advance, the Association's President-elect, acting as program chairman, decides on the topics for which sessions will be organized. This is done after consultation and comment, both volunteered and solicited, from a wide range of individuals. (A *Call for Papers* is published annually in the Notes section of the December issue of the *AER*.) The President-elect invites persons to organize these sessions. Each session organizer in turn invites several persons (usually two or three) to give papers on the theme of the session, and asks others to give comments on the papers. The program chairman decides at the time of organization which sessions are to be included in this volume. Space limitations restrict the number of printed sessions. This year we are printing 26 sessions, although a total of 107 sessions were sponsored, either solely by the American Economic Association or jointly with other allied societies. There is no standard practice with regard to the publication of comments and discussions, and each President-elect must decide how to allocate available space between invited papers and discussions. This year the President-elect invited the discussants on one session to submit written comments for publication. He also asked a discussant in another session to submit his comments in an independent paper.

The guidelines under which papers are published in the *Papers and Proceedings* differ from those governing regular issues of the *Review*. First, the length of papers is strictly controlled. Except in unusual circumstances they must be no more than twelve typescript pages in three-paper sessions and eighteen typescript pages in two-paper sessions. Second, papers are not subjected to any refereeing process. Third, their content and range of subject matter reflect the wishes of the President-elect to investigate and expose the current state of economic research and thinking. In most cases they are therefore exploratory and discursive, rather than formal presentations of original research.

While authors are encouraged to submit their manuscripts earlier, in practice most are submitted at the meeting itself, or in the four days following. Very rigid deadlines must be met and there is no time for communication with every author about editing changes made in order to improve content and style, and to satisfy space restrictions.

Rather than reject a paper because it is too long, every effort is made to reduce its length and, at the same time, preserve the main ideas. However, if such cuts do not seem feasible, we may ask the author to allow its consideration for publication in a regular issue of the *Review* subject to the usual refereeing process, or the author may be asked to withdraw the paper and submit it elsewhere. A paper is also rejected if, after reading it, we conclude that it is utterly without merit. This year we are pleased to report that no paper has been rejected on either ground.

JOHN G. RILEY
WILMA ST. JOHN

RICHARD T. ELY LECTURE

Self-Command in Practice, in Policy, and in a Theory of Rational Choice

By THOMAS C. SCHELLING*

An increasingly familiar occurrence for obstetricians is being asked by patients to withhold anesthesia during delivery. The physician often proposes that a facemask be put beside the patient who may inhale nitrous oxide as she needs it. But some determined patients ask that no such opportunity be provided: if gas is available they will use it, and they want not to be able to.

The request is interesting for decision theory, and raises questions of ethics, policy, and physician responsibility, even if the woman is merely making a mistake—if she simply does not know how painful labor will be and how glad she will be, even in retrospect, if the pain is relieved. But some women who make this request have had earlier deliveries during which they demanded anesthesia and received it. They are acquainted with the pain. They anticipate asking for relief. And they want it withheld when they do. They expect to regret afterwards any recourse to anesthesia.

This particular instance of attempted self-denial has features that are special but many that are common. The woman is, so far as we know, in good health physically and mentally. She anticipates a transient period when her usual values and preferences will be suspended or inaccessible. She has reasons for wanting to frustrate her own wishes at the critical time. She needs cooperation. She may ratify her choice afterward by expressing herself grateful that no anesthesia was offered, even when requested. There are ethical dilemmas and legal issues, and there is conflict, if, say, the husband disagrees with the physician in the delivery room about what his wife really wants.

I. Anticipatory Self-Command

This obstetrical example, though special in certain respects, is not a bad paradigm for the general anomaly of anticipatory self-command. That is the phenomenon that I want to discuss—that a person in evident possession of her faculties and knowing what she is talking about will rationally seek to prevent, to compel, or to alter her own later behavior—to restrict her own options in violation of what she knows will be her preference at the time the behavior is to take place. It is not a phenomenon that fits easily into a discipline concerned with rational decision, revealed preference, and optimization over time.

Attempting to overrule one's own preferences is certainly exceptional, as consumer behavior goes, but not so exceptional that anyone who reads this is unfamiliar with it. Let me remind you of some of those behaviors that share with obstetrical anesthesia the characteristic that a person may request now that a later request be denied. Please do not give me a cigarette when I ask for it, or dessert, or a second drink. Do not give me my car keys. Do not lend me money. Do not lend me a gun.

Besides denial there are interventions. Do not let me go back to sleep. Interrupt me if I get in an argument. Push me out of the plane when it's my turn to parachute. Don't let me go home drunk unless you can remove my children to a safe place. Blow the fuse if you catch me watching television. Make me get up and do my back exercises every morning.

*Lucius N. Littauer Professor of Political Economy, Harvard University, Cambridge, MA 02138. I am grateful to the Russell Sage Foundation and the Alfred P. Sloan Foundation for support and encouragement in this work.

Keep me moving if I am exhausted in the wilderness. Pump my stomach if you catch me overdosed with sleeping pills.

Then there is restructuring of incentives, often with somebody's help. Wagers serve this purpose, and are often used by people who share an interest in losing weight. Confessing something incriminating that can be revealed in the event of a lapse, or just making a ceremonial display of determination to exercise or to stay off cigarettes, can threaten oneself with shame.

Most of the tactics used to command one's own future performance probably do not depend on someone else's participation. I mentioned some that do, partly for comparison with the obstetrical example, partly because our experience with purely individual efforts is usually restricted to our own and we are unaware of the efforts of others unless a need for cooperation makes them visible. Further, the legal, ethical, and policy issues arise mainly when a second party is enlisted. And these are the cases that appear to call for a judgment about the ambivalent person's true interest—which set of preferences deserves our loyalty or sympathy.

The obstetrical case is rich in its ethics and legalities. To which patient is a physician obligated? The one asking for anesthesia or the one who asked that it be withheld? Can the physician enter a contract that will both protect against malpractice and compel compliance with the woman's earlier preferences? Do we like policies that make such contracts possible; do we like policies that make such contracts void?

Physicians, of course, are bound by a professional code as well as their personal ethics, and are subject to criminal and civil complaints. In the same way, our personal ethics are challenged when the drinking guest who entrusted us with his car keys wants them back, or snatches them and heads for his car. Our ethics are even challenged when he didn't ask but we know he intended not to drive himself home, he has a momentary alcoholic confidence in his driving ability, he will certainly thank us tomorrow if we disable his car, but he demands now that we let him alone.

Professional discussion of suicide indicates that anticipation of changing preferences is common. There are two symmetrical cases here. One is preventing suicide when a person has asked for protection against his own determination during periods when he unmistakably prefers to be dead. The other is the contrary, being begged to expedite someone's departure in the event of some ghastly condition, even if the condition is accompanied by such horror of dying that he will beg us to perpetuate that horror in violation of our earlier promise. There is also the person who elects death but cannot face the finality of bringing it about, and, like the parachutist who asks to be shoved out if he grips the door jam, implores our help in getting him over the brink.

Legal issues arise in some attempts to abdicate rights that are deemed to be inalienable. I cannot get a court injunction against my own smoking. I cannot contract with a skydiving pilot to push me out of the airplane. I cannot authorize my psychiatrist in advance to have me hospitalized against my wishes in circumstances that we have agreed on. I cannot contract with a fat farm to hold me against my will until I have lost some number of pounds; they have to let me out when I ask. (If we are clever we can arrange it; I go to a remote fat farm that requires a 24-hour notice to order a car, a notice that I can rescind during a moment's resurgent resolve to lose weight. I have heard that what keeps cruise ships from offering this kind of service is the inability to keep the crew from smuggling extra calories on board for the black market.)

An interesting issue is the ethics of prohibition—against, say, the display and sale of rich desserts in the faculty dining room, or against cigarette smoking in the workplace—not to keep others from overeating or smoking, as is usually the motivation behind prohibitions, but to keep ourselves from succumbing and to reduce the pain of temptation. There is a legal test in Massachusetts now of whether nicotine addiction is a protected species of handicap and a person has a right to relief through smoking in the workplace.

The most serious cases are those that involve, one way or another, actively or passively, taking your own life—one of your selves taking the life that you share. The law takes sides with the self that will not die. Someone who lives in perpetual terror of his own suicidal tendencies can welcome the law's sanctions against people whom he might, during a passing depression, beg to help with suicide. People for whom life has become unbearable but who cannot summon the resolve to end it have the law against them in their efforts to recruit accomplices. In December a California judge ruled against a quadriplegic woman who wished to die and asked the hospital's help in starving herself to death. The judge ordered forcefeeding, with the comment that "our society values life."

Besides legal issues there are regulatory policies. Nicotine chewing gum is being introduced as a prescription drug. The National Academy of Sciences has proposed that cigarettes low in tar and high in nicotine be developed to see whether people can better regulate their intake of tars, carbon monoxide, and other gasses if they can more readily satisfy their need for nicotine. And female hormones are being administered to violent male sex offenders who volunteer for treatment.

There are now remote monitors that can be attached to a parolee that will transmit encrypted messages at scheduled times through an attachment to the parolee's telephone to monitor whether he is abiding by a curfew. But he could voluntarily submit to surveillance by a friend, spouse, or other guardian; and I remind you of the electric-shock dog-training collars that can administer a deterrent to misbehavior. There is no technical difficulty in devising an unremovable blood-alcohol monitor that could activate a radio signal, or even administer a painful shock.

There are dangers. One can imagine a variety of self-restraining or self-compelling measures that could be used as conditions for employment, for election to office, for borrowing money, or for parole or probation, if it were known that one could incur an ostensibly voluntary enforceable commitment. The polygraph is a current example. Sterilization is another.

Many heroin addicts are alcoholics. Methadone is legally available for some heroin addicts; it replaces the need for heroin. Antabuse is legally available for alcoholics; it interacts with alcohol to produce extreme nausea, and precludes drinking. Methadone is attractive—at least in the absence of heroin—but antabuse is unattractive when alcohol is available. Some therapists provide the methadone only after the patient has taken the antabuse in the presence of the therapist.[1]

II. Self-Command and the Rational Consumer

How can we accommodate this phenomenon of strategic self-frustration in our model of the rational consumer? We can begin by asking whether there is a single phenomenon here, one that can be epitomized by addiction, appetite, or pain.

Adam Smith, by the way, included a chapter on self-command in his *Theory of Moral Sentiments*. He meant something different—courage, generosity, and other manly virtues. In my usage, self-command is what you may not need to employ if you already have enough of what Adam Smith meant by it. You don't need the skillful exercise of self-command to cope with shifting prefer-

[1] There is an "interaction effect" that sometimes has to be taken into account in judging the merits of voluntarily incurred coercion, or even involuntarily. Physicians who advise their cardiac and pulmonary patients about smoking, and psychiatrists who deal with hospitalized (incarcerated) heroin addicts, report a common phenomenon. Addicts suffer noticeably less withdrawal discomfort when in an establishment that has a reputation for absolute incorruptibility, unbribable guards and staff, and no underground market anywhere, compared with a hospital in which it is expected, rightly or wrongly, that appropriate effort and willingness to pay will produce relief. Cardiac and pulmonary patients who are told flatly that they must stop completely, at once, if they want to survive the year not only quit more frequently than patients merely advised to quit if they can, or, if they can't, to cut down or switch brands, but—this is the parallel to the heroin example—report surprisingly less withdrawal discomfort than those who succeed in quitting after getting the less absolute advice.

ences if you've already got your preferences under control. I cannot resist quoting a passage that I'm sure he'd like an opportunity to edit once more. "We esteem the man who supports pain and even torture with manhood and firmness; and we can have little regard for him who sinks under them, and abandons himself to useless outcries and womanish lamentations."

There is a quite heterogeneous array of types and circumstances and it will be useful to recall them. What they have in common is that they invite efforts at anticipatory self-command. Many of them are quite ordinary.

We can begin with behavior anticipated when one is fatigued, drowsy, drunk, or coming out of a sound sleep. Or for that matter asleep: people do misbehave in their sleep. They scratch; they remove dressings from wounds; they adopt postures not recommended by orthopedists. Wearing mittens to frustrate scratching or putting the alarm clock across the room are perfectly familiar techniques of self-command.

Quite different are acute thirst and hunger, panic, pain, and rage; some athletes drink water through straws to avoid gulping, and many people forego the advantages of a gun in the house for fear they'll use it.

There is captivation—books, puzzles, television, argument, fantasy—that engage a person against his earlier determination not to be so engaged. Keeping your mind from misbehaving on its own is somewhat different from keeping it from making wrong decisions; still, the mind that sneaks off into reverie without permission, or that won't stop chewing on some logical paradox, can be thought of as actually consuming—against orders.

There are phobias—reactions of admittedly unreasoning fear to heights, enclosures, crowds, audiences, blood, needles, reptiles, leeches, filth, and the dark. These, too, look sometimes like the mind misbehaving; several of them can be brought under some control by shutting one's eyes. It is not only pediatricians who suggest looking away when the knee has to be drained through a four-inch needle. I've seen many references to a phenomenon I experienced as a child—the dark is not so frightening if you shut your eyes, especially under the bedclothes.

There are compulsive personal habits involving faces and fingernails that are difficult to frustrate because we cannot take a trip and leave our cuticles behind.

Certain illnesses entail such protracted depression that, just as a person may attempt to make decisions now that he cannot change when he becomes aged, a person may put certain decisions beyond reach during an anticipated postoperative depression. It is not for nothing that we have the phrase, "a jaundiced view"; hepatitis does change one's outlook profoundly. Medication can change a person's values; self-administration of drugs, stimulants, and tranquilizers is used deliberately to alter one's effective preferences, and can have similar effects inadvertently. Alcohol makes some people brave when they need to be brave and some foolhardy when they can't afford to be. People for whom medicinally induced swings in mood are an unavoidable chronic way of life shouldn't be disqualified as the rational consumers that our theoretical assumptions are supposed to represent.

Some of those behaviors, like falling asleep, may not sound like consumer choices, possibly because we do not usually identify them with the marketplace, and some may not seem altogether voluntary. They do remind us that attempts to achieve self-comand are familiar, not necessarily abnormal, and when abnormal not uncommon.

There are many such behaviors that we have to acknowledge do look like consumer choice: smoking, drinking, overeating, procrastination, exercise, gambling, licit and illicit drugs, and shopping binges. And remember, I am speaking only of people who want to deny themselves later access to the foods, drugs, gambling, sexual opportunities, criminal companionship, or shopping splurges that constitute their own acknowledged problems in self-command. Anyone who is happily addicted to nicotine, benzedrine, valium, chocolate, heroin, or horse racing, and anyone unhappily addicted who would not elect the pains and deprivations of withdrawal, are not my subject. I am not concerned with whether cigarettes or rich desserts are bad for you, only with the fact that there are people who wish so badly to avoid them that, if they could, they would put those commodities beyond their own reach.

It is not an invariable characteristic of these activities that there is a unanimously identified good or bad behavior. Some dieters try to stay below a healthy body weight. Some people are annoyed at teetotalers, successful dieters, compulsive joggers, or people who never lose their tempers. And somebody who pleads for help in taking his own life, and alternately pleads not be be heeded on the occasions when he does, offers no easy choice as to who it is we should prefer to win the contest. The same is true of people who take steps to prevent their own defection from some religious faith.

While all of the cases I mentioned, from scratching to religious conversion, are within the subject of self-command, not all of them need to be recognized in a theory of rational decision. The person who prefers not to get out of bed we can consider just not all there; there are chemical inhibitors of brain activity that play a role in sleep, and until they have been metabolized away his brain is not working. His case may typify important decisions, but not the ones our theory is about. You can't make rational decisions when you're not rational, and you should rationally keep yourself from trying. Noisy alarms out of reach represent a rational choice.

What we can do is to append to our consumer a list of disqualifying circumstances in which his decisions are likely to be mistaken ones, and we make it the ordinary consumer's business, if he can't keep out of those circumstances, to take steps in advance to keep himself from making any decisions, or to arrange in advance to have his decisions disregarded. An important part of the consumer's task is then not merely household management but self-management—treating himself as though he were occasionally a servant who might misbehave. That way we separate the anomalous behavior from the rational; we take sides with whichever consumer self appeals to us as the authentic representation of values; and we can study the ways that the straight self and the wayward self interact strategically. We can adopt policies that, if they don't cause troubles elsewhere like interfering with civil liberties, help the consumer in his rational moments to control that other self and to keep important decisions from falling into the wrong hands.

But what about the person who, having given up cigarettes six months ago, succumbs after dinner to an irresistible urge to light a cigarette, who does so in apparent possession of his faculties, who six months earlier, or six hours, would have paid a price to ensure that cigarettes would be unavailable at the moment he changed his mind? If he were crazed with thirst or acutely suffering opiate withdrawal we could disqualify the decision: the mind is partly disconnected, a level of mind has taken over that is incapable of handling more than a couple of primitive dimensions of desire. But the person lighting that cigarette doesn't look as though he's bereft of his higher faculties.

The conclusion I come to is that this phenomenon of rational strategic interaction among alternating preferences is a significant part of most people's decisions and welfare and cannot be left out of our account of the consumer. We ignore too many important purposive behaviors if we insist on treating the consumer as having only values and preferences that are uniform over time, even short periods of time.

Just to establish the magnitude of the problem, consider cigarette smoking. There are thirty-five million Americans who have quit smoking. Most of them had to make at least three serious tries in order to quit. Of those thirty-five million, about five million are in danger of relapse, and two million will resume smoking and regret it. Most of those will try again, and three-quarters will fail on the next try. There are fifty-five million cigarette smokers, among whom some forty or forty-five million have tried to quit; nearly half have already tried three times or more, and some twenty million of those cigarette smokers made a serious try, and failed, within the past year. More than half of all young smokers, of both sexes, tried to quit within the past year and failed. A third of all young smokers have unsuccessfully tried three times or more. They know that smoking is dangerous, and we know that it is worth some years of their life expectancy. Smoking behavior alone is a major determinant of consumer welfare, one that a theory based on stable preferences and rational choice cannot illuminate without some modification; and smoking is only one such behavior.

There has been interesting work on how time preferences, as among future points in time, can change as time goes by—how one's preferred allocation of resources between the decade of the 1990's and the next decade after that can change between 1980 and 1990. I have in mind ideas associated with Robert Strotz (1956), Edmund Phelps and Robert Pollak (1968), Pollak (1968), and Jon Elster (1977, 1979). And we know the anecdote of the politically radical twenty-year-old whose conservative father infuriates him by putting a sum of money in trust that the son may use for political contributions only when he reaches the conservative age of forty. I propose we admit not only unidirectional changes over time, but changes back and forth at intervals of years, months, weeks, days, hours, or even minutes, changes that can entail bilateral as well as unilateral strategy.[2]

There are different ways to say what I'm describing. Two or more sets of values alternately replace each other; or an unchanging array of values is differentially accessible at different times, like different softwares that have different rules of search and comparison, access to different parts of the memory, different proclivities to exaggerate or to distort or to suppress. We know that the sight of a glistening bowl of peanuts can trigger unintended search and retrieval from memory, some of it subliminal, and even changes in the chemical environment of the brain. In common language, a person is not always his usual self; and without necessarily taking sides as between the self we consider more usual and the other one that occasionally gains command, we can say that it looks as if different selves took turns, each self wanting its own values to govern what the other self or selves will do by way of eating, drinking, getting tattooed, speaking its mind, or committing suicide.

III. Strategy and Tactics

From this point of view we can be quite straightforward in examining the strategies and tactics with which different selves compete for command. Here are some of the strategies I have in mind.[3]

Relinquish authority to somebody else: let him hold your car keys.

Commit or contract: order your lunch in advance.

Disable or remove yourself: throw your car keys into the darkness; make yourself sick.

Remove the mischievous resources: don't keep liquor, or sleeping pills, in the house; order a hotel room without television.

[2] An imaginative and comprehensive treatment of this subject, including comparisons with animal behavior, is George Ainslie (1975). An intriguing philosophical approach is Elster (1977, 1979). In economics there are attempts to fit self-control within the economics tradition and some outside that tradition. The best known effort to fit self-control within the economics tradition is George Stigler and Gary Becker (1977); their formulation denies the phenomenon I discuss. On the edge of traditional economics are C. C. von Weizsacker (1971) and Roger McCain (1979). Outside the tradition and viewing the consumer as complex rather than singular are Amartya Sen (1977), Gordon Winston (1980), Richard Thaler and H. M. Shefrin (1981), and Howard Margolis (1982). Winston, Thaler-Shefrin, and Margolis recognize a referee or superself, or planner-doer dichotomy, that I do not see; whether the difference is perception or methodology I am not sure. The most pertinent interdisciplinary work I know of by an economist is the brilliant small book by Tibor Scitovsky (1976). For related earlier work of mine, see my 1984 book.

[3] These strategies exclude "seek professional help," even "get a good book." There are therapies: some are based on fairly unified theories and some are quite eclectic. Good examples in print of the more eclectic are K. Daniel O'Leary and G. Terrence Wilson (1975) and David Watson and Roland Tharp (1981), intended for use as college textbooks, and Ray Hodgson and Peter Miller (1982), a serious work designed for popular use. Many of the strategies I mention are represented in books like these. A more focussed self-help book is Nathan Azrin and R. Gregory Nunn (1977), now unfortunately out of print; it deals mainly with "grooming" and other personal habits.

Submit to surveillance.

Incarcerate yourself. Have somebody drop you at a cheap motel without telephone or television and call for you after eight hours' work. (When George Steiner visited the home of Georg Lukacs he was astonished at how much work Lukacs, who was under political restraint, had recently published—shelves of work. Lukacs was amused and explained, "You want to know how one gets work done? House arrest, Steiner, house arrest!")

Arrange rewards and penalties. Charging yourself $100 payable to a political candidate you despise for any cigarette you smoke except on twenty-four hours' notice is a powerful deterrent to rationalizing that a single cigarette by itself can't do any harm.[4]

Reschedule your life: do your food shopping right after breakfast.

Watch out for precursors: if coffee, alcohol, or sweet desserts make a cigarette irresistible, maybe you can resist those complementary foods and drinks and avoid the cigarette.

Arrange delays: the crisis may pass before the time is up.

Use buddies and teams: exercise together, order each other's lunches.

Automate the behavior. The automation that I look forward to is a device implanted to monitor cerebral hemorrhage that, if the stroke is severe enough to indicate a hideous survival, kills the patient before anyone can intervene to remove it.

Finally, set yourself the kinds of rules that are enforceable. Use bright lines and clear definitions, qualitative rather than quantitative limits if possible. Arrange ceremonial beginnings. If procrastination is your problem, set piecemeal goals. Make very specific delay rules, requiring notice before relapse, with notice subject to withdrawal. Permit no exceptions.[5]

IV. Implications for Welfare Judgments

An unusual characteristic of these two selves, if you will permit me to call them selves, is that it is hard to get them to sit down together. They do not exist simultaneously. Compromises are limited, if not precluded, by the absence of any internal mediator. I suppose they might get separate lawyers or agree on an arbitrator. If the obstetrician with whom I began this lecture insists on taking the pain somewhat more seriously than his patient wanted him to, we would have an arbitrated compromise between the two selves.

For this reason we should expect outcomes that occasionally appear Pareto nonoptimal compared with the bargains they might like to strike:

Not keeping liquor or rich foods in the house, both selves suffering the detriment to their reputation as host;

Not keeping sleeping pills in the house, both selves suffering occasional insomnia;

Not keeping television in the house, both selves missing the morning news.

The simplicity with which we can analyze the strategy of self-command by recognizing the analogy with two selves comes at a price —a price in terms of what we value in our model of the consumer. When we identify a consumer attempting to exercise command over his own future behavior, to frustrate some of his own future preferences, we import into the individual a counterpart—I

[4] There is a cocaine addiction clinic in Denver that has used self-blackmail as part of its therapy. The patient may write a self-incriminating letter that is placed in a safe, to be delivered to the addressee if the patient, who is tested on a random schedule, is found to have used cocaine. An example would be a physician who writes to the State Board of Medical Examiners confessing that he has violated state law and professional ethics in the illicit use of cocaine and deserves to lose his license to practice medicine. It is handled quite formally and contractually, and serves not only as a powerful deterrent but as a ceremonial expression of determination.

[5] My back book prescribes exercises that are to be done faithfully every day. I am certain that some of them need to be done only two or three times a week. But the author knows that "two of three times a week" is not a schedule conducive to self-disciplines. My periodontist tells me that patients told to perform certain cleansing operations faithfully every day are pretty good at it, but told they can get along on two or three times a week relapse to two or three times every two or three weeks; he cannot then credibly insist they go back on the daily schedule.

think an almost exact counterpart—to interpersonal utility comparisons. Each self is a set of values; and though the selves share most of those values, on the particular issues on which they differ fundamentally there doesn't seem to be any way to compare their utility increments and to determine which behavior maximizes their collective utility.

I should remark here that it is only in talking with economists that I feel at all secure in using the terminology of "selves." Philosophers and psychiatrists have their own definitions of the self, and legal scholars may resist the concept of the multiple self when it seems to raise questions about which "self" committed the crime or signed the contract, and whether the self on trial is the wrong one and we must wait for the "other" to materialize before trial, sentence, or incarceration. It is only in economics that the individual is modelled as a coherent set of preferences and certain cognitive facilities; and though economists are free to deny the phenomenon I'm discussing, if they recognize the phenomenon I think they have little difficulty with the language of alternative selves.

What about that woman who denies herself anesthesia, pleads for it during delivery, and denies it again at the next delivery? What about the person who drops by parachute with survival gear into the wilderness to go a month without smoking, drinking, overeating or sleeping late as he beats his way back to civilization, cursing all the way the self that jumped, then pleased with himself when the ordeal is over? Is there a way to formulate the question, did the individual maximize utility? Or can we only argue that one of the selves enhanced its own utility at the expense of the other? When we ask the mother who an hour ago was frantic with pain whether she is glad the anesthesia was denied her, I expect her to answer yes. But I don't see what that proves. If we ask her while she is in pain, we'll get another answer.

As a boy I saw a movie about Admiral Byrd's first Antarctic expedition and was impressed that as a boy he had gone outdoors in shirtsleeves to toughen himself against the cold. I decided to toughen myself by removing one blanket from my bed. That decision to go to bed one blanket short was made by a warm boy; another boy awoke cold in the night, too cold to go look for a blanket, cursing the boy who removed the blanket and swearing to return it tomorrow. But the next bedtime it was the warm boy again, dreaming of Antarctica, who got to make the decision, and he always did it again. I still don't know whether, if those Antarctic dreams had come true, I'd have been better able to withstand the cold and both boys would have been glad that the command structure gave the decision to the boy who, feeling no pain himself, could inflict it on the other.

The person who can't get himself up in the morning I said was not quite all there. Why does that count against him? Apparently because he cannot fully appreciate what it will be like to be late to work. But does the self who sets the alarm, and arranges with a tennis partner to roll him out of bed, fully appreciate the discomfort of getting out of bed? My answer is yes. But notice: I am not in bed. I lecture only when I am awake, and the self that might prefer to stay in bed goes unrepresented.

In another respect I am not impartial. I have my own stakes in the way people behave. For my comfort and convenience I prefer that people act civilized, drive carefully, and not lose their tempers when I am around or beat their wives and children. I like them to get their work done. Now that I don't smoke, I prefer people near me not to. As long as we have laws against drug abuse it would be easier all around if people didn't get hooked on something that makes them break the law. In the language of economics, these behaviors generate externalities and make us interested parties. Even if I believe that some poor inhibited creature's true self emerges only when he is drunk enough to admit that he despises his wife and children and gets satisfaction out of scaring them to death, I have my own reasons for cooperating with that repressed and inhibited self that petitions me to keep him sober if I can, to restrain him if he's drunk, or to keep his wife and children safely away from him.

Consider the person who pleads in the night for the termination of an unbearable existence and expresses relief at midday that his gloomy night broodings were not taken seriously, who explains away the nighttime self in hopes of discrediting it, and pleads again for termination the next night. Should we look for the authentic self? Maybe the nighttime self is in physical or mental agony and the daytime self has a short memory. Maybe the daytime self lives in terror of death and is condemned to perpetuate its terror by frantically staying alive, suppressing both memory and anticipation of the more tangible horrors of the night. Or the nighttime self is perhaps overreacting to nocturnal gloom and depressed metabolism, trapped in a nightmare that it does not realize ends at dawn.

The question, which is the authentic one, may define the problem wrong. Both selves can be authentic. Like Siamese twins that live or die together but do not share pain, one pleads for life and the other for death—contradictory but inseparable pleas. If one of the twins sleeps when the other is awake, they are like the two selves that alternate between night and day. The problem seems to be distributive, not one of identification.

A few years ago I saw again the original Moby Dick, an early talkie in black and white. There was a scene—not in the book—of Ahab in the water losing his leg, and immediately afterward below deck under a blanket, eating an apple with three of the crew. The blacksmith enters with a hot iron to cauterize the stump. Ahab begs not to be burned. The crewmen hold him down as he spews out the apple in a scream, and steam rises where the iron is tormenting his leg. The movie resumes with Ahab out of pain and apparently glad to be alive. There is no sign that he took disciplinary action against the blacksmith or the men who held him while he was tortured.

When I first began contemplating this episode I thought it an incontestable case of the utility gain from denying freedom of choice and ignoring revealed preference. I wondered whether Ahab might have instructed the blacksmith that in the event of a ghastly wound to any member of the crew it was the blacksmith's responsibility to heat an iron and burn the wound, even if the wounded man were Captain Ahab. However much he implores us now not to burn his leg, Ahab will surely thank us afterwards. But now I wonder what that proves.

If one of *you* were to be burned so that *I* might live I would probably thank the people who did it. If you burn *me* so that I may live I'll thank you, afterward, but that is because I'll be feeling no pain and not anticipating any when I thank you. Suppose I were to be burned and Ahab in the next room needed to be cauterized too. Would you, while holding me down in disregard of my plea, ask my expert advice on whether to burn Ahab, and his advice on whether to burn me?

How do we know whether an hour of extreme pain is more than life is worth? Alternatively, how do we know whether an hour of extreme pain is more than death is worth?[6] The conclusion that I reach is that I do not know, not for you and not for me.

I do feel sure that if I wanted in such circumstances to endure the pain I would have to rely on people who were tough enough in spirit to hold me down, or at least to tie me down. And if any violation of the Captain's express orders constituted mutiny punishable by death, you would have to gag Ahab to keep him from screaming "don't" and thus condemning himself to a fatal infection. (Still, if the Captain himself presides over the trial of the mutineers who held him when he shouted "stop," they will be in no danger of his wrath; so, anticipating acquital with thanks, they may as well hold him down.)

I have found, in conversations about Ahab's plight, that people like me approve of his being burned against his express wishes, not merely burned despite his involuntary

[6] Many discussions of ambivalence toward suicide, especially for the wretchedly or terminally ill, suggest a comparison with the case of Ahab. The ambivalence appears less an alternation between preferences for life and for death than a preference for death and a horror of dying. Death is the permanent state; dying is the act of getting there, and it can be awesome, terrifying, gruesome, and possibly painful. Ahab can enjoy life—minus a leg—only by undergoing a brief horrifying event, just as the permanent relief of death can be obtained only by undergoing what may be a brief and horrifying event, especially if the healing professions will not help or are not allowed to.

screams and thrashings but against his horrified begging before he went out of his mind with pain. I interpret that to mean that people like me prefer a regime in which we ourselves would be held and burned even if we asked not be be. Yet our willingness to consider the need to be held against our will is an acknowledgement that, being certainly no braver than Ahab, we would in the event react as he did. That could mean that, at a position remote in time or in likelihood from the event we are better able to appreciate the relative merits of pain and death. But when I examine my own attitude, I usually find the contrary. If I try to imagine my way into Ahab's dilemma I find myself becoming so obsessed with immediate pain compared with immediate death that I begin agreeing with Ahab.[7]

If there is any wisdom in my current choice, which is to be held and burned if I am ever in Ahab's situation, it is the wisdom of choosing sides without fully acquainting myself with their merits. What I avoid is identifying myself with that person who may be burned, even though I know that it could be I. In the same way afterwards, I shall thank you because I do not much identify with the historical I who was burned in the recent past. But I shall know then that if I had to do it again I would prefer death. It is hard for two selves that do not simultaneously exist to compare their pains, joys, and frustrations.

In exploring this problem of identity I have been tantalized by some imaginary experiments: imagine being offered a chance to earn a substantial sum, say an amount equal to a year's income, for undergoing an exceedingly painful episode that would have no physical aftereffects. Upon hearing what the pain is like, you refuse; maybe you'd undergo it for twice that sum. The experimenter is embarrassed; anticipating your favorable response, he has already initiated the experiment with you, perhaps through something you drank. You suffer the pain and are confirmed in your original judgement that you wouldn't do it for a year's income. When the pain is over and you've recovered from the shock, you receive the money. Question: when you see the experimenter on the sidewalk as you test-drive your new Porsche, are you glad he made that hideous mistake?

A second experiment: some anesthetics block transmission of the nervous impulses that constitute pain; others have the characteristic that the patient responds to the pain as if feeling it fully but has utterly no recollection afterwards. One of these is sodium pentothal. In my imaginary experiment we wish to distinguish the effects of the drug from the effects of the unremembered pain, and we want a healthy control subject in parallel with some painful operations that will be performed with the help of this drug. For a handsome fee you will be knocked out for an hour or two, allowed to sleep it off, then tested before you go home. You do this regularly, and one afternoon you walk into the lab a little early and find the experimenters viewing some videotape. On the screen is an experimental subject writhing, and though the audio is turned down the shrieks are unmistakably those of a person in pain. When the pain stops the victim pleads, "Don't ever do that again. Please."

The person is you.

Do you care?

Do you walk into your booth, lie on the couch, and hold out your arm for today's injection?

Should I let you?

REFERENCES

Ainslie, George, "Specious Reward: A Behavioral Theory of Impulsiveness and Impulse Control," *Psychological Bulletin*, July 1975, *82*, 463–96.

Azrin, Nathan H. and Nunn, R. Gregory, *Habit Control in a Day*, New York: Simon and Shuster, 1977.

Elster, Jon, "Ulysses and the Sirens: A The-

[7] I find it difficult to predict my choice if I were in a situation comparable to Ahab's but with a choice whether to initiate my remaining life with the agonizing episode or to postpone the pain until later. It always seems to me that anyone able to elect the pain at all would be tempted to take it now. Pain in the future may be discounted, but pain past is discounted more. Faced with an episode of frightening pain, people often do try to get it over and done with.

ory of Imperfect Rationality," *Social Science Information*, 1977, *41*, 469–526.

———, *Ulysses and the Sirens*, Cambridge: Cambridge University Press, 1979.

Hodgson, Ray, and Miller, Peter, *Self-Watching —Addictions, Habits, Compulsions: What to Do About Them*, New York: Facts on File, 1982.

McCain, Roger A., "Reflections on the Cultivation of Tastes," *Journal of Cultural Economics*, June 1979, *3*, 30–52.

Margolis, Howard, *Selfishness, Altruism, and Rationality*, Cambridge: Cambridge University Press, 1982.

O'Leary, K. Daniel and Wilson, G. Terrence, *Behavior Theory: Application and Outcome*, Englewood Cliffs: Prentice Hall, 1975.

Phelps, Edmund S. and Pollak, R. A., "On Second-Best National Saving and Game-Theoretic Equilibrium Growth," *Review of Economic Studies*, April 1968, *35*, 185–99.

Pollak, R. A., "Consistent Planning," *Review of Economic Studies*, April 1968, *35*, 201–08.

Schelling, Thomas C., "The Intimate Contest for Self-Command," in his *Choice and Consequence*, Cambridge: Harvard University Press, 1984, 57–82.

———, "Ethics, Law, and the Exercise of Self-Command," in *Choice and Consequence*, 83–112.

———, "The Mind as a Consuming Organ," in *Choice and Consequence*, 328–46.

Scitovsky, Tibor, *The Joyless Economy: An Inquiry into Human Satisfaction and Consumer Dissatisfaction*, New York: Oxford University Press, 1976.

Sen, Amartya K., "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," *Philosophy and Public Affairs*, Summer 1977, *6*, 317–45.

Smith, Adam, "Of Self-Command," in *The Theory of Moral Sentiments*, Section III, Part VI, 1759.

Stigler, George J. and Becker, Gary S., "De Gustibus Non Est Disputandum," *American Economic Review*, March 1977, *67*, 76–90.

Strotz, Robert H., "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, No. 3, 1956, *23*, 165–80.

Thaler, Richard H. and Shefrin, H. M., "An Economic Theory of Self-Control," *Journal of Political Economy*, April 1981, *89*, 392–406.

von Weizsäcker, C. C., "Notes on Endogenous Changes of Tastes," *Journal of Economic Theory*, December 1971, *3*, 345–72.

Watson, David L. and Tharp, Roland G., *Self-Directed Behavior: Self-Modification For Personal Adjustment*, Monterey: Brooks/Cole Publishing, 1981.

Winston, Gordon C., "Addiction and Backsliding," *Journal of Economic Behavior and Organization*, 1980, *1*, 295–324.

Improving the Teaching of Economics: Achievements and Aspirations

By G. L. BACH AND ALLEN C. KELLEY*

Apparently somewhere between two-thirds and three-fourths of all economists make a significant part of their living by teaching. Yet only a minuscule part of their professional training is focused on how to become a good teacher. It is not clear that our teaching performance is much better now than it was twenty-five or fifty years ago.

The Committee on Economic Education (CEE) of the American Economic Association (AEA) has since the 1950's operated under a broad and challenging mandate—to help improve the teaching in college and university economics. The CEE played a major role in a big push to improve teaching at all levels during the 1960's. Activities included a National Task Force on Economic Education, comprised of six of the nation's most distinguished economists who were asked to suggest the economic essentials that each citizen-to-be should have in our democratic, largely market-directed economy; a year-long CBS and PBS economics course, including work for credit at many colleges and universities; a nationwide program of summer workshops for teachers; and a major push on research on the effectiveness of different kinds of teaching.

Several of these projects were established on a continuing basis. Others served their purposes and faded away. Thus, the CEE of the 1970's inherited a large agenda of work-in-progress. But much as the CEE of the 1950–60's accomplished, there was much more to be done to improve the teaching of economics at all levels, and, in 1973, Bach as out-going chairman spelled out a challenge, "An Agenda for Improving the Teaching of Economics" for the next decade.

The new committee developed a wide-ranging program of continuing and new projects. The committee's activities have been "project oriented." These projects, with approximately one million dollars of foundation grants over the decade, will be discussed here around three broad themes: training economic educators, developing economic-education research, and establishment of the *Journal of Economic Education* (*JEE*). We then conclude with a description of a number of projects being considered now by the committee, and some which we as individuals would like to see seriously considered.

I. The Training of Economic Educators

Achievements. The 1973 committee agenda placed heavy emphasis on producing more well-trained teachers in economics. The university incentive system was, and still is, stacked against students becoming distinguished teachers. Few major economics departments offered any training in how to teach. In a painful majority of cases, the new TA was simply put in the classroom with a copy of the textbook and good wishes. If the TA was so bad the students complained bitterly, he heard about it. But the positive rewards were for outstanding research, few for good teaching. Given these facts, the committee has repeatedly urged that departments take positive steps to improve the teaching abilities of new economists.

*Frank E. Buck Professor of Economics and Public Policy, Emeritus, Stanford University, Stanford, CA 94305, and James B. Duke Professor of Economics, Duke University, Durham, N.C. 27706, respectively. Bach was chairman of the CEE from 1964 to 1977; Kelley served from 1977 to 1983. The views expressed herein are solely our own, and do not necessarily represent those of past or present CEE members.

In the 1970's, the committee established a major experimental Teacher Training Program (TTP) for new Ph.Ds and young teachers of economics. The thrust of the TTP was to develop a teacher training program, or model, that was fundamental yet flexible enough to be adaptable to many courses and school settings. Planning and Policy Committees were established to define the problem, to pull together what we know about good teaching and what techniques are most effective under different conditions, and to establish actual training programs for economists.[1]

A decade later, the project has resulted in several conspicuous "outputs":

1) A *Resource Manual* was prepared, covering course planning, classroom behavior, lecturing, cases, leading discussion and other teaching styles, and examinations. Video tapes are used for teacher self-evaluation. The *Manual* provides the most reliable, validated information available on these questions. It also includes descriptions of actual training programs that range from one-day "crash" courses for TAs to full-semester courses given for graduate credit.

2) Several workshops have been held in which teams of graduate students and faculty members from leading Ph.D-granting universities have met for one to two weeks to learn how to use the *Manual* and supporting materials, and how to institute training programs at their own schools. These multi-school workshops have thus far been held at Indiana University-Bloomington, University of Wisconsin-Madison, University of North Carolina-Chapel Hill, Harvard University, and University of Colorado-Boulder. Over 300 graduate students and faculty members, representing sixty colleges and universities, have participated.

3) In the past decade, one-school training programs have been conducted at twenty-five institutions, and approximately ten of these have instituted their own TTP on a regular basis. Roughly 1,000 graduate students have participated in these programs. Presently, training programs for new Ph.Ds and teachers are well established at Harvard, Indiana, Minnesota, North Carolina, and Wisconsin. Numerous others have mandated some formal program of teacher training.

Understandably, there has been considerable doubt that we know much about how students learn best, or what is the best way to teach. But psychologists have completed a vast amount of research on these issues. There is a large amount of empirical evidence, much of it carefully statistically tested, on how humans learn under different conditions. There is certainly no one "best" way to teach everything under all conditions. But there are some ways that are pretty surely better than others for inducing student learning under different conditions.

Aspirations. We judge that the teacher training program has been a substantial success. The number of schools involved, and graduate students and young teachers trained, has exceeded our expectations. But there is still much to be done. Expansion of programs is surely not the only way to improve teaching in economics. But it could do a lot. Over 1,000 new Ph.Ds and established faculty members have been helped by a TTP in the past five years. The CEE's aspiration is to at least double this number in the next five years. We note, incidentally, that the most effective results occurred when (a) junior and senior faculty members of the same school have participated in TTP workshops; this provides status and reinforcement for each other when such formal training in teaching may face cynicism on their home campus; and (b) where the TTP is an established, accepted part of faculty development. The CEE plans to continue support for new TTPs around the country. It is our aspiration that these programs become well-established parts of all major Ph.D programs, and of faculty development programs for other

[1] The Planning Committee included Arthur Welsh (chair), W. Lee Hansen (chairman of the CEE), Allen C. Kelly, Darrell Lewis, and Phillip Saunders; the Policy Committee was G. L. Bach (chair), James Buchanan, Rendigs Fels, R. A. Gordon, Walter W. Heller, and George Stigler. Financial support was initially obtained from the Alfred P. Sloan Foundation and subsequently from the Lilly Endowment, Exxon Educational Foundation, JM Foundation, City Corporation, Joint Council on Economic Education, and several participating universities. Approximately one-half million dollars has been spent on this project.

young teachers. We especially hope to establish regional TTPs, where smaller, less well-financed schools can send their younger teachers at moderate cost. We are also experimenting with small "teams" of experienced TTP leaders that might help at the financially more limited schools.

II. Research on Economic Education

Achievements. The Committee on Economic Education has emphasized the promotion of research on economic education for three reasons. First, adding to the stock of knowledge about what determines success in teaching economics represents a direct assault on our primary goal of improving teaching. Second, the theoretical perspectives and empirical tools of the economics discipline appear peculiarly germane to the study of a process (learning) which involves the transformation of inputs into outputs. Third, professional respect and credibility in economics is largely related to scholarly activity. If economic education as a "field" of inquiry is to be fully established, it too must devote a substantial portion of its emphasis to research.

Expanding a research base is a slow process, but, happily, we can report some significant achievements. The quantity and quality of research on economic education and learning have steadily improved. Moreover, the results from this research are beginning to accumulate into useful, well-established generalizations about teaching and learning economics. "Hard" research involving explicit hypothesis testing has been replacing the "soft" casual empiricism of the past. Sessions reporting on economic education research are now commonplace at national and regional professional meetings. The research has benefited enormously from the development of a standardized test, the Test of Understanding College Economics (TUCE), originally sponsored by the Committee in 1968, and revised in 1981. This test is the most widely used measure of economic "outputs" in the various articles appearing the *JEE* and the AEA *Papers and Proceedings*.

A comprehensive review of economic education research since World War II was conducted by John Siegfried and Rendigs Fels. Based on a review of 179 pieces of research, they concluded:

> A cumulative literature on economics education has now developed.... The quality of the research done so far varies widely, but dramatic improvement has occurred in recent years; much more of the current research is first rate than was the case 15 years ago. ... The field [economic education] is rapidly becoming respectable, and the research findings can be useful to college economics teaching. [1979, p. 959]

Most research on economic education has focused on the principles course. In an effort to broaden this emphasis, the CEE has backed a census of all economics departments to determine the nature of the economics major in the United States. This ambitious project, led by John Siegfried, has just been completed, and the last of several published papers presenting this new, large data base is being given in this session today.

Aspirations. The advancement of knowledge is a slow process. We have now established a modest research base, publishing outlets, and the professional respectability required to attract more scholars into economic education research. The various "projects" of the committee—the *Journal of Economic Education*, the Test of Understanding College Economics, the sessions at the annual AEA meetings and the published papers in the *Proceedings*, the sessions sponsored at various professional society meetings, and the study of the major—have significantly encouraged the advancement of knowledge in economic education. During the next decade the CEE hopes to strengthen these activities, and attention will be focused especially on the quality of the effort.

III. The Journal of Economic Education (*JEE*)

Achievements. Founded in 1969 by the CEE and the Joint Council on Economic Education (JCEE), the *JEE* has primarily published research reports on teaching. Most of these studies have been on the principles course, almost all have involved empirical analysis, and a large share have used controlled-experimental techniques to ferret out the impacts of alternative teaching technologies. Under the founding editorship of Henry Villard, the *JEE* has had notable success, measured not only by its 2,000-plus subscriptions, but also on less quantifiable dimensions: 1) it has stimulated research on economic education, and served as a notable outlet for this research; 2) it has helped establish the credibility of economic education within the economics profession; and 3) it has encouraged professionals to identify with the economic education movement.

One decade after its founding, the CEE and the JCEE undertook a retrospective review of the *JEE* to document its achievements and to explore plans for the 1980's. The result of that review was a decision to expand significantly the scope of the *JEE*. It was agreed that a strong research orientation should be maintained. But the goals of the *JEE* should be expanded to include more interests of teachers.

To organize this effort, the editorship was rotated to Donald Paden (University of Illinois). In turn, he has recruited four Associate Editors, each taking responsibility for one section of the "new" *Journal*. The new focus is conveyed by its new sections.

1) *Research in Economic Education* (William Becker, Jr., Indiana University). This section publishes original research on economic education—the evaluation of teaching methods; student learning, attitudes and interests, and teaching materials and processes. These continue as the lead sections of the *JEE*.

2) *Content Articles in Economics* (Kalman Goldberg, Bradley University). This section features articles on substantive issues, especially when the article throws light on new developments in economics, or provides an especially good example of analysis of current problems.

3) *The Teaching of Economics* (Donald Paden, University of Illinois). This section publishes articles on the teaching of economics and the teaching profession. It includes essays on great teachers, how to cope with the knowledge explosion, economic literacy, and other related issues in education.

4) *Innovation in Economic Instruction* (Karl Case, Wellesley College). This section publishes relatively brief articles, notes, and communications, dealing with *new* pedagogical developments, hardware, or teaching materials, and novel ways of treating traditional subject matter.

5) *Professional Information* (Robin Bartlett, Dennison University). This section includes reports on economics enrollments, the economics major, the labor market for economists, salaries, job placements, the status of women and minorities in teaching, and related topics.

Aspirations. The ambitious goal of the new *JEE* is to become indispensable to teachers of economics. The editors welcome the support of the profession, both in readership and in providing contributions to the *JEE*. (Communications may be sent to Professor Donald W. Paden, Managing Editor, David Kinley Hall, 1407 W. Gregory Drive, Urbana, IL 60801, or to the appropriate associate editor.)

IV. Promising Projects: The Future

The activities stated above will continue to occupy a substantial part of the CEE's attention for some time to come. But a number of other ideas have been suggested and considered by the committee. We, as individuals, suggest that there are still others worth considering.

A. *Use of Computers in Teaching Economics*

The so-called "computer revolution," nose-diving computing costs, and the increase in student "computer literacy" suggest that computers may play a much bigger role in teaching economics tomorrow. Yet computer technology has not caught the imagination or commitment of most economic educators. We lag many other fields, notably the sciences and engineering.

Some schools have projects well along, trying out the use of computers in both undergraduate micro- and macroeconomics courses. Several schools have used simple-to-complex macro policy games. A scattering of individuals have done some interesting work in the micro areas—but only a scattering.

The CEE tentatively proposes to bring together a small study group of "experts" that will provide a report on the key teaching areas where the computer promises productive use, and perhaps those where it does not look very hopeful. Ways in which software needs can be effectively developed need consideration. The extent to which this development should be largely "market based" or sponsored by nonprofit institutions, especially the government, remains an open question. The committee is inclined toward primary reliance on the market, but there are serious externality problems because of the great difficulty in software developers' claiming for themselves the benefits from the use of the software. Software is often hard to protect from being reproduced. Quite possibly, the society in one way or another should take responsibility for developing some kinds of teaching and research-oriented software, much as it finances research.

Leaving aside this problem, the informal study group's job would be to explore the promise of computers in teaching economics, and to provide information to the profession on what is being done, how effectively computers have been used to date, and the outlook for the future. The study group's report would presumably be published in the *JEE*, *AER*, or *JEL*.

B. *Use of Radio in Teaching Economics*

Another, more established, technology that might be used more effectively to teach economics is radio, in particular, the more than 1,000 college radio stations and the Public Broadcasting System. Economics taught over TV and the radio has been used many times, including a year-long course offered in the early 1960's under the policy guidance of a committee appointed by this profession—with a regular daily listening audience over CBS and PBS of over one million persons.

Use of short radio "spots," offered by college campus broadcasting stations and possibly by PBS stations, to get across particularly intriguing and currently important points in economics, seems to several members of the committee a worthwhile possibility. One approach is the use of one- or two-minute spots to raise interesting problems with economics, to suggest possible answers to these, and, especially, to draw in the listener on a "What would you do about it?" basis. Such topics might include both local and national issues—for example the distribution of tickets to athletic events or parking spaces, markets for college graduates, the pricing of campus food services, the impacts of inflation, and the like. Answers to the problem might or might not be suggested at the time of the program. Spot-type approaches have been used with some promise in other fields, notably law.

Such spots would be interest-arousers—a chance to stir up the public's and students' interest in more formal economics. The programs would need to be highly professional, and made with the heavy participation of experts in communication. The committee is now investigating the costs and possible usefulness of a small "pilot" experiment with such programs.

C. *Learning Theory and Research as a Foundation for Teaching Economics*

If we want to teach more effectively, surely one first step should be to find out what is known about human learning, the conditions under which it occurs, the different ways people acquire knowledge, and so on. Yet almost no economists lay such a foundation for teaching economics. Indeed, often the attitude toward theory and the evidence produced by psychologists and others studying human learning is indifference or a sneering one. Implicitly, we learn about teaching by crude empiricism—trying something ourselves, or perhaps listening to another economist who is alleged to be a good teacher. How much accumulated knowledge about learning is applicable to economics is not clear. But not to make any use of the research seems to some members of this committee inefficient, somewhat provincial, intellectual behavior. We urge that the gains from building a bridge between learning theorists and other experts on human learning, and economists who are interested in teaching more effectively, promise real rewards. Some of us believe this is priority research that may offer "practical" knowledge about teaching and learning that we can use to improve our teaching.

We are, of course, aware that there is no widely accepted "general theory" of human learning. The experts in the field, however, know a great deal about the conditions under which different kinds of learning take place most effectively, and those unfavorable to learning.

Progress in learning from the experts on learning will, at best, be slow and perhaps difficult. But in the long run it will provide a firmer and deeper foundation for good teaching.

D. *Increasing the Supply of Women as Teachers of Economics*

It is widely agreed that the supply of women economists should be increased. Some of the reasons for the small representation of women among professional economists and in college and university teaching are well known. One striking result of numerous studies is that women do less well in undergraduate economics courses than do men who appear similar on other scores. Numerous possible causes have been advanced, but the answer is not clear. Some members think the CEE should put this question on the urgent research menu, since undergraduate performance is an important control over admission to graduate school, and thence to college-university teaching jobs.

Others argue that it is probably necessary to mount an ambitious, concerted research attack on the role of gender in economics. Why the "gender gap" in grades and in many economics jobs? They argue that understanding the gender gap is just part of the general problem of understanding the supply of men and women trained economists.

The AEA already has a major standing committee on the place of women in the profession. It is not clear just whose responsibility lies where. We want to cooperate and some members of CEE are exploring what channels of research seem most promising in this area.[2]

E. *Use of Cases in Teaching*

Teaching "cases" are widely used in business, law, and other professional schools, and many other fields, including political science. The CEE generally supports leaving publication of teaching materials to commercial channels. However, the potential teaching benefits from the use of cases seem to us so large that a reasonable presumption can be stated for this committee to encourage development of such materials in economics.

There is no one simple, generally accepted definition of teaching cases. Business school cases, perhaps most widely known by economists, are usually detailed statements of a business situation, including the external environment, in which the student has to make a decision, often acting as president or some other decision-making official. Business school cases tend to be essentially exercises in management decision making, generally without the use of any formal theory. However, powerful uses can be made of cases as examples of the problems of *applying* theories to real world situations, and it is this second approach that we find most promising for economics. Or cases can be used to explore how effectively the theory's assumptions map the "real world" situation. Other uses are possible.

Many standard economics textbooks now include cases, but some of these show a limited understanding by the authors of what cases are and how they can be used effectively. Merely to clip out a news story and use it for a discussion of issues in the course is not necessarily to gain the advantages of case analysis, partly because material is often not available in the clipping to develop information needed for analysis.

F. *Cooperation with Schools of Education*

Most regular college and university courses in economics are taught by instructors who

[2] The committee is aware that a similar case can be made for special investigation of the relative position of members of minority groups on grades and on jobs held. Since the committee is limited by available time and funds, most of the members at least tentatively prefer going ahead with the study of the gender gap problems, not least because the data available on academic performance of women as against similarly placed men is much more complete than that on minority groups in comparison with nonminorities.

have had at least some professional training in economics. But the great majority of grade- and high-school teachers teaching economics (often in other social studies courses) have had little or no course work in economics themselves. Very few states require that social studies teachers take any courses in economics as a condition of obtaining a social studies teacher's certificate—though now over half of all states mandate the teaching of either a full course or part of a course about economics (often about the "private enterprise system") in the schools. The schools of education primarily responsible for college training of such teachers often have no economists on their faculties, and require little or no work in economics outside the school of education. The status of schools of education dropped sharply during the 1960's and 1970's, reflecting widespread criticism of the quality and lack of "toughness" of the educational programs in the schools. But interest has been shifting back toward relying on schools of education to improve the undergraduate programs received by future teachers. Thus we suggest this may be a time for serious attempts to work cooperatively with teachers and administrators in education—where having more good economics available to students could make a real difference in the quality of economics teaching received by students in the grade and high schools in the years to come. How to come to grips with the problem is not easy.

V. The Joint Council on Economic Education

Over the years the CEE has been enhanced by our association with the Joint Council on Economic Education. The JCEE is a nonprofit, nonpartisan, educational organization to encourage, improve, coordinate, and service economic education from kindergarten through college. It has a large network, including 50 state Councils, and 242 Centers operating on college and university campuses. The work of the CEE is facilitated by the fund-raising support of the JCEE, the administration and vesting of CEE foundation grants, and the staff at the Council assigned to higher education. The JCEE staff has made large contributions to developing and carrying out CEE projects. In turn, the Council has benefited from the direct participation of CEE members in the appraisal and formulation of curricula and materials in economics for the nation's schools; and the impact of different CEE programs to increase the supply of well-trained economists teaching in the precollege and elementary college levels. This mutually supportive working relationship is of great value to both organizations.

Information about many of the joint activities of the CEE and JCEE (for example, the Teacher Training Program, and the Test of Understanding College Economics) may be obtained by writing the Director, College and University Programs, Joint Council on Economic Education, 2 Park Avenue, New York, NY 10016.

VI. Conclusion

The CEE's mandate is large; our resources are limited. We invite suggestions of activities we should undertake or encourage, and comments on projects currently under way or proposed. Moreover, we actively solicit expressions of interest to participate in the projects of the Committee. We serve the American Economic Association, and are eager to be responsive to interests of our colleagues in our collective effort to advance economic education in the United States. Information on the chair and members of the committee, and on its programs, is provided annually in the committee reports section of the *American Economic Review Papers and Proceedings*.

REFERENCES

Bach, G. L., "An Agenda for Improving the Teaching of Economics," *American Economic Review and Proceedings*, May 1973, *63*, 303–08.

Siegfried, John J. and Fels, Rendigs, "Research on Teaching College Economics: A Survey," *Journal of Economic Literature*, September 1979, *17*, 923–69.

A Profile of Senior Economics Majors in the United States

By JOHN J. SIEGFRIED AND JENNIE E. RAYMOND*

The American Economic Association's (AEA) Committee on Economic Education and the Joint Council on Economic Education (JCEE) have sponsored a comprehensive study of the economics major in the United States.[1] This article reports results of an April 1981 survey of 1,080 senior majors in economics at 48 colleges and universities who graduated in spring 1981. An earlier report presented the results from a survey of 546 departments that offered a bachelor's degree with a major in economics.[2]

Information in this article may help departments to identify strengths and weaknesses in their students and programs, and may be useful to the economics profession generally by revealing the educational goals and career plans of our students. The data collected in our survey are available for further research on economics education.[3]

The student questionnaire required approximately 15 minutes to complete. Respondents were invited to report their names and permanent addresses to facilitate a follow-up study, but they were also urged to complete the questionnaire even if they preferred to remain anonymous. Seventy-one percent of the respondents provided name and address. These students were contacted again in May 1983 and asked to verify their educational and career experience; 398 responded.

The questionnaire was administered by a faculty member at each of 50 colleges and universities. The sample institutions were chosen to provide a representative distribution of respondents on six criteria: region of the country, total institution enrollment, exclusive gender of students, predominant race of students, private or public control, and degree level in economics (whether they offer graduate degrees). The percentage of respondents in each category is reported in column (3) of Table 1. Table 1 also reports the distribution of economics degrees conferred across these categories based on National Center for Education Statistics (NCES) data for 1978 (the most recent year for which data were then available).

Sample schools were selected so that their enrollments generated approximately the same distribution of students across the criteria as did 1978 degrees conferred. Variation in response rates at individual schools caused deviations between the actual sample distribution and the goal. Our sample includes too many students from the Southeast, the smaller schools, and private institutions. These are the only statistically significant (*Chi*-square test) deviations between the distributions in column (1) and column (3) of Table 1. Because the survey is not based on a random sample, we generally avoid formal statistical testing.

The sample appears to be substantially above average academically, with an estimated average combined SAT score of 1,216 compared with the nationwide average for graduating seniors (in all disciplines) in 1981

*Professor of Economics and Graduate Student in Economics, respectively, Vanderbilt University, Nashville, TN 37235. This project received enormous encouragement and aid from Allen Kelley, W. Lee Hansen, Arthur Welsh, Rendigs Fels, John Soper, James Wilkinson, Kaye D. Evans and Katherine M. McElroy. Phillip Saunders, Michael Salemi, Marianne Ferber, Jum Nunnally, Robert Highsmith, John Sumansky, Stephen Buckles, J. S. Butler advised us on the design of the study. We are most grateful to the 48 professors who distributed and collected the survey and the 1,080 students who completed it.

[1] The Sloan Foundation provided primary financial support for this project. Supplementary support was received from the JCEE, AEA, and the University Research Council of Vanderbilt University.

[2] See Siegfried and James Wilkinson. That study reports on the number of economics major programs in the United States, the characteristics of their students and faculty, requirements for the major, availability of special educational opportunities, course enrollments, and the effect of curriculum design on the choice of majors by students.

[3] For information on the public use data tapes produced by this project, contact the Joint Council on Economic Education, 1212 Avenue of the Americas, New York, NY 10036.

TABLE 1—SAMPLE STATISTICS FOR SPRING 1981 SURVEY OF SENIOR ECONOMICS MAJORS[a]

	Percentage Distribution of		
	Economics BAs conferred 1978[b] (1)	1980–81 Senior Economics Majors[c] (2)	Respondents (3)
Region			
Northeast	45.4	40.6	40.9
Midwest	20.7	32.7	24.0
Southeast	12.6	11.5	24.3
West	21.2	15.2	10.8
Enrollment			
0–999	4.3	2.1	3.2
1000–9999	44.2	46.3	62.6
10000+	51.2	51.6	34.3
Sex			
Male	1.3	1.4	2.8
Female	2.8	5.0	7.8
Coed	95.9	93.6	89.4
Control			
Public	52.2	53.4	35.6
Private	47.1	46.6	64.4
Offering Level			
No grad. econ.	16.5	16.8	21.3
Grad. program in economics	83.5	83.2	78.7
Predominate Race			
White	98.4	97.6	99.4
Other	1.6	2.4	0.6

[a] Fifty colleges and universities were initially sampled; 48 schools provided at least one usable response.

[b] Based on National Center for Education Statistics data tape.

[c] Based on fall 1980 questionnaire to departments (see Siegfried and Wilkinson, 1982).

of 890. In selecting the sample institutions, we relied for administering the questionnaire on faculty members who were known to the project staff, members of the AEA Committee on Economic Education, or the JCEE staff. Most of these faculty are located at academically stronger colleges and universities. The liberal arts colleges in the sample all were classified in the higher academic quality category by the Carnegie Commission. Furthermore, a response bias can be expected in favor of the more successful students, who would be more anxious to tell us about a favorable experience in their major.

TABLE 2—GENERAL STUDENT CHARACTERISTICS OF SENIOR ECONOMICS MAJORS BY TYPE OF INSTITUTION

	Carnegie Code Classification			
	All Schools	*R&D*[a]	*COMP*[b]	*LA*[c]
Sample				
Schools	48	20	11	17
Students	1,080	564	195	321
Transfer Students[d]	18.3	19.4	25.1	12.2
Women[d]	36.5	32.6	32.8	45.9
Foreign[d]	3.3	3.2	4.6	2.8
Nonwhite[d]	4.7	3.8	7.8	4.4
Private HS[d]	24.1	25.4	17.1	26.0
Took HS Econ.[d]	29.8	27.2	33.0	32.5
SAT				
Verbal	589	590	590	586
Quantitative	627	624	629	631

[a] *R&D* is research and doctorate institutions; schools awarding at least 10 Ph.D.s annually.

[b] *COMP* is comprehensive institutions; those with a minimum of 1,000 students, fewer than 10 annual Ph.D.s and at least one professional program plus liberal arts.

[c] *LA* is liberal arts colleges.

[d] Shown in percent.

To aid schools in comparing their economics majors to those in similar institutions, we adopted the Carnegie Code Classification developed by the Carnegie Commission on Higher Education (1973). The number of sample institutions and students in each category are reported in Table 2.

I. General Student Characteristics

General student characteristics are reported in Table 2. Thirty-six percent of the students are women (in contrast to the 29 percent of 1981 economics graduates reported by the 546 departments; see Siegfried and Wilkinson, p. 128). There are fewer transfer students and more women among senior economics majors at liberal arts colleges. Nonwhite economics majors are twice as abundant at comprehensive institutions. The SAT scores are remarkably similar across the various types of colleges.

The sample students come from well-educated families. Forty-seven percent of their mothers and 68 percent of their fathers hold a bachelor's degree. Advanced degrees are held by 15 percent of their mothers and 38 percent of their fathers, and 40 percent of the students have at least one parent working in business management.

Twenty-four percent of the students attended a private high school. While in high school, either public or private, 30 percent took an economics course. The high school economics course appears to be substantially more than a course in personal finance. Seventy-seven percent of the students reported that they covered supply and demand analysis in their high school course; 48 percent covered basic macroeconomic issues; and 19 percent covered the concepts of elasticity and marginal analysis. Only about 25 percent of the students reported covering personal finance issues such as filling out tax forms in their high school economics course.

II. The Decision to Major in Economics

All students did not decide to major in economics at the same stage in their educational career. Twelve percent report that they had decided to major in economics before arriving at college; 19 percent decided in their freshman year; 46 percent in their sophomore year; 20 percent in their junior year; and 3 percent in their senior year. Students who did better on the quantitative portion of the SAT examination and students whose parents hold advanced degrees in economics (1.8 percent of the sample) elected to major in economics sooner than other students.

Table 3 reports the importance students attach to various reasons for majoring in economics. Interest in the subject, improved prospects for employment, practical knowledge for decision making, and preparation for professional school are the most popular rationales. The advice of parents and friends, departments' research reputations, and grading standards are the least common reasons. Even at research and doctorate universities, only 5 percent of the students admit to considering the research reputation of the department as being very important in their decision to major in economics.

An earlier report from this project (Siegfried and James Wilkinson) examined the influence of certain curriculum design factors on students' choice of major. Apparently only one major curriculum factor affects the number of students who major in economics—if there is an undergraduate business administration degree available at the same institution the number of economics majors declines drastically.

TABLE 3—IMPORTANCE OF REASONS FOR MAJORING IN ECONOMICS

	Students Responding "Very Important"[a]			
Reason	All Schools ($n=1080$)	*R&D* ($n=564$)	*COMP* ($n=195$)	*LA* ($n=321$)
Interest in Subject	72.1	70.6	75.4	73.2
Better Prospects for Employment	45.2	42.0	45.6	50.5
Practical Knowledge for Decision Making	35.2	35.3	37.9	33.6
Preparation for Professional School	33.2	35.3	28.7	32.4
Favorable Impression of Faculty	31.3	30.0	31.8	33.6
Challenged by Analytical Method	30.6	29.3	31.3	33.3
Better Prospects for Higher Earnings	30.0	28.5	28.2	33.6
Complement to Another Major	23.8	28.4	21.5	17.1
Teaching Reputation of Dept.	20.3	16.7	16.9	28.7
Program Flexibility	15.1	20.9	10.8	7.5
Performance Better in Econ. Courses	11.7	13.5	11.8	8.4
Preparation for Graduate Study	7.4	6.9	10.8	6.5
Parents' Advice or Example	6.9	6.2	7.2	7.8
Advice of Friends	4.1	4.4	3.1	4.0
Research Reputation of Dept.	3.8	4.8	3.1	2.5
Can Earn Higher Grades in Econ.	2.4	2.8	2.1	1.9

[a] Shown in percent. The response options were irrelevant, somewhat important, and very important.

III. The Students' Experience

The grade point average (GPA) of the students in the sample is 3.23 for their economics courses and 3.18 overall (on a four-point basis). The women in the sample have an overall GPA about 0.1 higher than the men, but both men and women have equivalent GPAs in economics. There is no difference between U.S. and foreign students on either the overall or economics GPA. There is no difference between whites and nonwhites in the overall GPA, but a 0.2 advantage for whites appears in the economics GPA.

p14420

TABLE 4—MEAN SCORES ON STANDARDIZED GRADUATE AND PROFESSIONAL SCHOOL ADMISSIONS EXAMINATIONS, 1980–81

Exam	Nationwide Mean Score	Senior Economics Major Sample Mean Score All Schools	*R&D*	*COMP*	*LA*
GRE:					
Verb	469	571 (123)	581	550	593
Quant.	533	652 (122)	659	639	673
Anal.	498	644 (105)	651	628	650
Econ.	606	624 (142)	661	658	613
GMAT	468	585 (306)	588	565	591
LSAT	550	643 (246)	645	660	635

Note: Sample size is in parentheses for all institution samples.

The typical senior major has written 4.5 term papers of 5 or more pages in economics courses. The standard deviation of 3.5 term papers indicates considerable variation in students' writing experiences.

A large number of students had taken the standardized admission examinations for graduate and professional schools by the time the survey was completed. Mean scores are reported in Table 4. The extremely high academic talents of the sample students are again evident.

A high proportion of students took the economics Graduate Record Examination GRE (13 percent) because several institutions in our sample require all graduating seniors to take it as a senior comprehensive examination. Thus the sample of students taking the economics GRE includes some who have no intention of applying for admission to graduate economics degree programs (only 7 percent of the sample acknowledged plans to seek a graduate degree in economics), which explains the small advantage the sample has over the nationwide average on the economics GRE relative to the sample's large positive increment over nationwide averages on the other admissions tests.

TABLE 5—THE RELATIONSHIP BETWEEN STUDENT CHARACTERISTICS AND ECONOMICS PERFORMANCE[a]

	Dependent Variables		
Independent Variable	GPA[a] (1)	GRE[b] (2)	GRE (3)
Constant	1.455	29.326	204.918
Overall GPA			
(3.18; 0.44)	–	83.212[d]	–
[3.19; 0.48]		(6.78)	
SAT: Verbal			
(590; 83)	0.001[d]	0.293[d]	0.390[d]
[577; 82]	(6.21)	(3.88)	(4.36)
SAT: Quant			
(628; 81)	0.002[d]	0.135[c]	0.254[d]
[635; 75]	(8.00)	(1.72)	(2.74)
Sex (1 = female; 0 = male)			
(0.37; 0.48)	0.008	−33.413[d]	−28.449[c]
[0.23; 0.43]	(0.24)	(−2.39)	(−1.69)
Race (1 = nonwhite; 0 = white)			
(0.04; 0.20)	−0.120	−38.244	−62.688[c]
[0.04; 0.19]	(−1.53)	(−1.42)	(−1.95)
HS Econ. (1 = yes; 0 = no)			
(0.30; 0.46)	0.008	3.511	18.866
[0.34; 0.48]	(0.28)	(0.31)	(1.40)
Number of Papers of ≥ 5 pages			
(4.77; 3.57)	0.002	5.968[d]	4.929[c]
[4.66; 2.51]	(0.33)	(2.75)	(1.89)
Institution Intercept Shift Variables	45 unreported coefficients		
Coefficient of determination corrected for degrees of freedom)	.221	.635	.471
Number of observations (students)	897	128	128
Number of Institutions	46	22	22

Note: t-ratios are shown in parentheses under the reported coefficients; the respective mean and standard deviation for the samples are reported under the description of the independent variables, with the $n = 128$ sample in brackets.

[a] Dependent variable mean = 3.22; standard deviation = 0.49.

[b] Dependent variable mean = 626; standard deviation = 84.

[c] = significance at .05 level.

[d] = significance at .01 level.

IV. The Determinants of Student Performance in Economics

We estimated a standard regression "model" of student performance in economics courses. Complete sets of data are available for students from 46 different schools using economics GPA and for 22 schools using the economics GRE as a measure of student performance. As explanatory variables we include the usual SAT scores, sex and race binary variables, as well as an indicator of whether the student had taken a high school economics course, and the number of college economics papers the student had been required to write.

The results are reported in Table 5. The first regression shows variations in the economics GPA, and the second and third regressions show variations in student performance on the economics GRE. Column (2) includes a measure of the student's overall GPA; equations (1) and (3) omit this variable. Equation (2) tells us how important the various independent variables are for performance in economics relative to overall performance. It is useful for addressing the question: what determines whether students will do well in economics without regard to whether they might do well in other disciplines also.[4] Equations (1) and (3) tell us how important each independent variable is for performance in economics acting through its significance *for general academic performance as well as* its effect on *relative performance* in economics.

An intercept shift variable (binary variable) is included for each separate institution to control for possible systematic differences in the levels of GPAs or general talent of the students across schools.[5] The estimated coefficients for these variables are not reported.

So far as general performance in economics is concerned (equations (1) and (3)), both quantitative and verbal skills seem to matter. Women and nonwhites appear to have a disadvantage in performing well on the economics GRE, but are equivalent to men and whites, respectively, in GPAs. High school economics courses do not appear to be much help to students. If a high school economics course gives a college student an advantage, it does not appear to last throughout the four-year college period. Writing college economics papers does not appear to make much difference for one's economics GPA, but is associated with higher scores on the economics GRE. This difference might occur because courses in which papers are required are graded tougher than others, or it may be that students with higher GPAs elect courses with term papers assigned, and these students already are doing sufficiently well in their grades that experience writing papers does not help them improve their grades very much. Writing three additional term papers, which is slightly more than the standard deviation of number of term papers, is associated with about 15 additional points on the economics GRE, a substantial increment.

The results are comparable for identifying those factors that give students a *relative* advantage in economics (equation (2)). Both verbal and quantitative skills seem to matter. This is consistent with the evidence that, in general, verbal skills are more important for students' performance in the introductory economics course (Siegfried and Rendigs Fels, 1979, p. 937) and quantitative skills are more important for students' performance in graduate work in economics (W. Lee Hansen, 1971, p. 51). Our results are between those of the introductory course and graduate program results, but then our measures of student performance are for college seniors, who are between introductory economics and graduate work in economics.

Women appear at a relative and absolute disadvantage in performance in college economics (equations (2) and (3)). This is consistent with most of the research evidence (Siegfried, pp. 7–8). However, they do not appear at a disadvantage in GPA achievement.

Whites seem to score better than nonwhites, although the difference is statistically significant only for the equation measuring absolute performance on the economics GRE (equation (3)).

V. Career and Future Education Plans

Students' plans to continue their formal education are reported in Table 6. Eighty-seven percent of the students surveyed intended to continue their education beyond the baccalaureate degree in at least a year or two, although about 70 percent of the stu-

[4] Estimates for a parallel equation to predict economics GPA that includes overall GPA are not reported because the disturbance term for the overall GPA is correlated with the disturbance term for the economics GPA, which would bias the resulting coefficients. An instrumental variables technique cannot be used to avoid this problem because instruments that affect overall GPA are bound to also influence the economics GPA.

[5] No adjustment is made, however, for the possibility that the various independent variables may have different effects on performance at different schools.

TABLE 6—PLANS OF SENIOR ECONOMICS MAJORS TO CONTINUE THEIR FORMAL EDUCATION[a]

	All Schools	*R&D*	*COMP*	*LA*
No. of Students	1071	557	194	320
Plan to continue formal education:				
No plans	13.5	13.8	12.9	12.8
Immediately	33.7	36.4	34.0	27.8
In a year or two	53.7	49.7	53.1	59.4
Plan to attend				
Graduate School	26.2	24.2	36.8	23.1
Professional school	73.8	75.8	63.2	76.9
Graduate and professional school students planning to continue in economics	15.3	14.7	22.2	12.2

[a] Shown in percent.

TABLE 7—PLANS OF SENIOR ECONOMICS MAJORS TO SEEK FULL-TIME EMPLOYMENT IN YEAR AFTER GRADUATION

	All Schools	*R&D*	*COMP*	*LA*
No. of Students:	1069	559	191	319
Will not seek full-time employment	29.2	32.2	23.6	27.3
Will seek permanent full-time employment	45.2	45.4	48.7	42.6
Will seek temporary full-time employment	25.6	22.4	27.7	30.1

dents intended to seek employment immediately after graduation (see Table 7). There is a surprising similarity in career plans of students at different types of institutions.

Of the 398 students in the May 1983 follow-up survey, only 45 percent reported that they had *actually* enrolled in a post-baccalaureate degree program within two years of their graduation. But 83 percent of those in the follow-up survey had either continued their schooling *or* still planned to do so soon after May 1983.

Of the students who planned to attend professional school, 61 percent sought a business degree, 25 percent planned to pursue a law degree, and 9 percent sought both a business and a law degree. No other professional field was selected by more than 1 percent of those planning to attend professional schools. Of the 144 students in the follow-up survey who actually attended professional schools within two years of graduation, 49 percent were in business school and 47 percent were in law school.

Of the students who planned to enroll in graduate schools, 58 percent expected to major in economics, with the remainder widely scattered among other disciplines. Of the 40 in the follow-up survey who actually attended graduate school within two years, 22 selected economics as their field of specialization; 13 were enrolled in an economics Ph.D. program and 9 were in an economics M.A. program. Twenty-four additional students still planned to commence graduate work in economics. The fraction of senior economics majors who actually did or still planned to enroll in a Ph.D. program in economics is between 3.5 and 9.5 percent.

The most common occupational goals of those students seeking full-time employment were general management (33 percent), sales and marketing (12 percent), analyst (10 percent), researcher (6 percent), and economist (4 percent). Banking and finance was the most popular industry choice (34 percent), followed by "business in general" (8 percent), the computer industry (6 percent), and the insurance industry (4 percent). In actual experience, however, only 17 percent had attained a full-time position in general management two years after graduation; 20 percent described their position two years after graduation as an analyst, 14 percent were in sales and marketing, 10 percent were in intern programs, 8 percent were clerical workers, 5 percent were doing manual labor, and 4 percent were in data processing. Only 2 of the 289 respondents described their employment as an "economist."

The students expected to earn a median of $17,000 in their first year of full-time employment (regardless of when that occurs) and, in a world of no inflation, predicted their median earnings would be $54,900 twenty-five years later. This implies an expected real annual growth rate in earnings of 5.0 percent. The actual real annual growth rate in earnings for the cohort of 20–24 year-olds starting in 1950 has been 4.6 percent for men and 2.8 percent for women. Weighting the growth rates by the fraction of men and women in our survey yields a historical real annual growth rate of 4.0. The people in our sample are apparently optimistic about the future growth rate in real earnings.

REFERENCES

Hansen, W. Lee, "Prediction of Graduate Performance in Economics," *Journal of Economic Education*, No. 1, 1971, *3*, 49–53.

Siegfried, John J., "Male-Female Differences in Economic Education: A Survey," *Journal of Economic Education*, No. 2, 1979, *10*, 1–11.

______ **and Fels, Rendigs,** "Research on Teaching College Economics: A Survey," *Journal of Economic Literature*, September 1979, *16*, 923–69.

______ **and Wilkinson, James T.,** "The Economics Curriculum in the United States: 1980," *American Economic Review Proceedings*, May 1982, *72*, 125–38.

Carnegie Commission on Higher Education, *A Classification of Institutions of Higher Education*, Berkeley, 1973.

National Center for Education Statistics, Data tape: "Degrees and Other Formal Awards Conferred Between July 1, 1977 and June 30, 1978," HEGGIS, XIII, Department of Health, Education, Welfare, Education Division, Washington, 1978.

ECONOMIC EFFECTS OF THE RISE IN ENERGY PRICES: WHAT HAVE WE LEARNED IN TEN YEARS?

The Role of Energy in Productivity Growth

By DALE W. JORGENSON*

The objective of this paper is to analyze the role of energy in the growth of productivity. The special significance of energy in economic growth was first established in the classic study, *Energy and the American Economy 1850–1975*, by Sam Schurr and his associates (1960) at Resources for the Future. For the period from 1920 to 1955, Schurr noted that energy intensity of production had fallen while both labor and total factor productivity were rising.[1] The simultaneous decline in energy intensity and labor intensity of production ruled out the possibility of explaining the growth of productivity solely on the basis of substitution of less expensive energy for more expensive labor. Since the quantity of both energy and labor inputs required for a given level of output had been reduced, technical change is also a critical explanatory factor.

An alternative explanation for growth of output with declining energy and labor intensity required an examination of the character of technical change. This examination was motivated by the fact that from 1920 to 1955 the utilization of electricity had expanded by a factor of more than ten, while consumption of all other forms of energy only doubled. The two key features of technical change during this period were that the thermal efficiency of conversion of fuels into electricity increased by a factor of three and that "the unusual characteristics of electricity had made it possible to perform tasks in altogether different ways than if the fuels had to be used directly" (Schurr, 1983, p. 3).

Schurr (1982) has recently extended the analysis of *Energy and the American Economy* through 1981. For the period through 1969 his conclusions are as follows: "Although the inverse relationship between total factor productivity and energy intensity virtually disappeared during the 1953–1969 period, it is still noteworthy that high rates of improvement in total factor productivity were essentially not associated with increases in energy intensity" (1982, p. 6).

Schurr (1982) has analyzed the experience of the U.S. economy in the aftermath of the oil embargo of 1973. He points out that energy intensity of production has fallen steadily since 1973, and that the rate of decline accelerated sharply after the second oil price shock in 1979, following the Iranian revolution. He goes on to point out that:

> While energy productivity has been improving at a very high rate during the past decade, the overall productivity efficiency side of the story has been highly unfavorable, and has become a matter of great concern. The post-1979 years that witnessed a new high in the rate of growth of national energy productivity also saw a decline in productive efficiency with a *fall* in total factor productivity of about 0.3 percent per year between 1979 and 1981.
>
> [1982, p. 10]

I. Econometric Models

In order to assess the role of energy in stimulating productivity growth, it will be necessary to go behind trends in energy utili-

*Professor of Economics, Harvard University, Cambridge, MA 02138. This research was supported by the Electric Power Research Institute under RP 1152-6. Able research assistance was provided by Kun-Young Yun. Very useful comments on an earlier draft were made by Ernst Berndt, John Kendrick, and Sam Schurr. Responsibility for opinions expressed or for any deficiencies in the research are my own.

[1] See Schurr and Netschert et al. (1960). This summary is based on Schurr (1983).

zation and productivity. For this purpose I employ an econometric model of sectoral productivity growth of individual industrial sectors in the United States. In order to assess the significance of changing forms of energy, inputs are divided in each sector among capital, labor, electricity, nonelectrical energy, and materials. The econometric model encompasses substitution among productive inputs in response to changes in relative prices. This model also determines the growth of sectoral productivity as a function of relative prices.

For each industry my model of production will be based on a sectoral price function that summarizes both possibilities for substitution among inputs and patterns of technical change. Each price function gives the price of output of the corresponding industrial sector as a function of the prices of capital, labor, electricity, nonelectrical energy, and materials inputs and time, where time represents the level of technology in the sector.[2] Obviously an increase in the price of one of the inputs, holding the prices of the other inputs and the level of technology constant, necessitates an increase in the price of output. Similarly, if productivity of a sector improves and the prices of all inputs into the sector remain the same, the price of output must fall. Price functions summarize these and other relationships among the prices of output, capital, labor, electricity, nonelectrical energy, and materials inputs, and the level of technology.

The sectoral price functions provide a complete model of production patterns for each sector, incorporating both substitution among inputs in response to changes in relative prices and technical change in the use of inputs to produce output. To characterize both substitution and technical change, it is useful to express the model in an alternative and equivalent form. First, we can express the shares of each of the five inputs—capital, labor, electricity, nonelectrical energy, and materials—in the value of output as functions of the prices of these inputs and time, again representing the level of technology.[3] Second, we can add to these five equations for the value shares an equation that determines productivity growth as a function of the prices of all five inputs and time. This equation is my econometric model of sectoral productivity growth.[4]

Like any econometric model, the relationships determining the value shares of capital, labor, electricity, nonelectrical energy, and materials inputs and the rate of productivity growth involve unknown parameters that must be estimated from data for the individual industries. Included among these unknown parameters are biases of productivity growth that indicate the effect of change in the level of technology on the value shares of each of the five inputs. For example, the bias of productivity growth for electricity gives the change in the share of electricity in the value of output in response to changes in the level of technology, represented by time. We say that productivity growth is electricity using if the bias of productivity growth for electricity is positive. Similarly, we say that productivity growth is electricity saving if the bias of productivity growth for electricity is negative.

I have pointed out that this econometric model for each industrial sector of the U.S. economy includes an equation giving the rate of sectoral productivity growth as a function of the prices of the five inputs and time. The biases of technical change with respect to each of the five inputs appear as the coefficients of time, representing the level of technology, in the five equations for the value shares of all five inputs. The biases also appear as coefficients of the prices in the equation for the negative of sectoral produc-

[2] The price function was introduced by Paul Samuelson (1953).

[3] The sectoral price functions are based on the translog price function introduced by Laurits Christensen, myself, and Lawrence Lau (1971, 1973). The translog price function was first employed at the sectoral level by Ernst Berndt and myself (1973) and by Berndt and David Wood (1975). References to sectoral production studies incorporating energy and materials inputs are given by Berndt and Wood (1979).

[4] This model of sectoral productivity growth is based on that of myself and Barbara Fraumeni (1981). A useful survey of studies of energy prices and productivity growth is given by Berndt (1982).

tivity growth. This feature of our econometric model makes it possible to use information about changes in the value shares with time and information about changes in the rate of sectoral productivity growth with prices in determining estimates of the biases of technical change.

The dual role of the bias of productivity growth—expressing the impact of a change in technology on the value share of an input and the impact of a change in the price of that input—is the key to an assessment of the role of electrification in productivity growth. Historical evidence suggests that much of the innovation in the twentieth century is electricity using. Innovation increases the share of electricity in the value of output for a given set of input prices, including the price of electricity. Entirely different evidence, analyzed by Schurr and his associates, has linked the reduction in the cost of electricity resulting from increased thermal efficiency in electricity generation to enhanced productivity growth. Within my econometric model these two pieces of evidence are consistent with the hypothesis that the bias of productivity growth for electricity is positive.

II. The Role of Energy

A classification of industries by patterns of the biases of technical change is given in Table 1. The pattern that occurs with the greatest frequency is capital using, labor using, electricity using, nonelectrical energy using, and materials saving technical change. This pattern occurs for eight of the thirty-five industries included in my study. For this pattern the rate of technical change decreases with the prices of capital, labor, electricity, and nonelectrical energy inputs, and increases with the price of materials input. The pattern that occurs next most frequently is capital saving, labor using, electricity using, nonelectrical energy using, and materials saving technical change. This pattern occurs for five industries. For this pattern the rate of technical change decreases with the prices of labor, electricity, and nonelectrical energy inputs, and increases with the prices of capital and materials inputs. These two patterns of technical change differ only in the role of the price of capital input.

TABLE 1—CLASSIFICATION OF INDUSTRIES BY BIASES OF TECHNICAL CHANGE

Pattern of Biases[a]					
Capital	Labor	Electricity	Nonelectrical Energy	Materials	Industries
U	U	U	U	S	Tobacco, textiles, apparel, lumber and wood, printing publishing, fabricated metal, motor vehicles, transportation
U	S	U	U	U	Electrical machinery
U	U	U	S	S	Metal mining, services
U	U	S	U	S	Nonmetallic mining, miscellaneous manufacturing, government enterprises
U	S	U	S	U	Construction
U	U	U	S	U	Coal mining, trade
U	S	S	U	S	Agriculture, crude petroleum and natural gas, petroleum refining
S	U	U	U	U	Food, paper
S	U	U	U	S	Rubber, leather, instruments, gas utilities, finance, insurance and real estate
S	U	S	U	U	Chemicals
S	S	U	U	U	Transportation equipment and ordnance, communications
S	U	S	U	S	Stone, clay and glass, machinery
S	U	U	S	S	Primary metals
S	S	U	U	S	Electric utilities
S	S	S	S	U	Furniture

[a] S = saving; U = using.

I have found the technical change is electricity using for twenty-three of the thirty-five industries included in the study. The first and most important conclusion is that electrification plays a very important role in productivity growth. A decline in the price of electricity stimulates technical change in twenty-three of the thirty-five industries and dampens productivity growth in only twelve. Alternatively and equivalently, it can be said that technical change results in an increase in the share of electricity input in the value of output, holding the relative prices of all inputs constant, in twenty-three of the thirty-five industries. Technical change results in a decrease in the share of electricity input in only twelve of these industries.

The utilization of nonelectrical energy can also be examined. My findings are that technical change is nonelectrical energy using for twenty-eight of the thirty-five industries included in the study, and energy saving for seven of these industries. The second conclusion is that greater utilization of nonelectrical energy plays an even more significant role in productivity growth than electrification. A decline in the price of nonelectrical energy stimulates technical change in twenty-eight of the thirty-five industries and dampens productivity growth in only seven. Alternatively, one can say that technical change results in an increase in the share of nonelectrical energy input in the value of output in twenty-eight of the thirty-five industries and results in a decrease in the nonelectrical energy input share in only seven. I conclude that greater utilization of nonelectrical energy has a significant role in productivity growth for an even wider range of industries than electrification.

III. Conclusion

This paper has analyzed the role of electrification and the utilization of nonelectrical energy in productivity growth employing an econometric model of production and technical change. Within the framework provided by the model, I can offer a tentative explanation of the disparate trends in energy intensity and productivity growth. These trends first drew the attention of Schurr and his associates to the special role of electrification. Over the period 1920–53, energy intensity of production was falling while productivity was rising. While the fall in real prices of electricity and nonelectrical energy resulted in the substitution of energy inputs for other inputs, especially for labor input, these price trends also generated sufficient growth in output per unit of energy input that the energy intensity of production actually fell. This explanation is completely in accord with the explanation advanced by Schurr and his associates.

Between 1953 and 1973 energy intensity was stable, while productivity continued to grow. During this period, real energy prices continued to fall, but at slower rates than during the period 1920–53. As before, the fall in real prices of electricity and nonelectrical energy resulted in the substitution of energy input for other inputs. But this was almost precisely offset by the growth in output per unit of energy input, leaving the energy intensity of production unchanged. Finally, real energy prices began to rise in the early 1970's, increasing dramatically after the first oil shock in 1973 and again after the second oil shock in 1979. These price trends resulted in the substitution of capital, labor, and materials inputs for inputs of electricity and nonelectrical energy, thereby reducing energy intensity of production. At the same time, the energy price trends contributed to a marked slowdown in productivity growth.

Although much research remains to be done before we obtain a complete understanding of the role of energy utilization in productivity growth, it is important to emphasize the important progress toward achieving this goal. I have analyzed the character of technical change in a wide range of industries covering the whole of the U.S. economy, and have confronted hypotheses advanced in earlier research by Schurr and his associates with new empirical evidence. The data support the hypothesis that electrification and productivity growth are interrelated. Somewhat surprisingly, the data also show that the utilization of nonelectrical energy and productivity growth are even more strongly interrelated.

REFERENCES

Berndt, Ernst R., "Energy Price Increases and the Productivity Slowdown in United States Manufacturing," in *The Decline in Productivity Growth*, Boston: Federal Reserve Bank of Boston, 1982, 60–89.

______ **and Jorgenson, Dale W.,** "Production Structures," in Dale W. Jorgenson and

Hendrik S. Houthakker, eds., *U.S. Energy Resources and Economic Growth*, Washington: Energy Policy Project, 1973, ch. 3.

______ **and Wood, David O.**, "Technology, Prices, and the Derived Demand for Energy," *Review of Economics and Statistics*, August 1975, *56*, 259–68.

______ **and** ______, "Engineering and Econometric Interpretations of Energy-Capital Complementarity," *American Economic Review*, September 1979, *69*, 342–54.

Christensen, Laurits R., Jorgenson, Dale W. and Lau, Lawrence J., "Conjugate Duality and the Transcendental Logarithmic Production Function," *Econometrica*, July 1971, *39*, 255–56.

______, ______, **and** ______, "Transcendental Logarithmic Production Frontiers," *Review of Economics and Statistics*, February 1973, *55*, 28–45.

Jorgenson, Dale W. and Fraumeni, Barbara M., "Relative Prices and Technical Change," in E. R. Berndt and B. Field, eds., *Modeling and Measuring Natural Resource Substitution*, Cambridge: M.I.T. Press, 1981, 17–47.

Samuelson, Paul A., "Prices of Factors and Goods in General Equilibrium," *Review of Economic Studies*, October 1953, *21*, 1–20.

Schurr, Sam, "Energy Efficiency and Productive Efficiency: Some Thoughts Based on American Experience," *Energy Journal*, No. *3*, 1982, *3*, 3–14.

______, "Energy Efficiency and Economic Efficiency: An Historical Perspective," in Sam Schurr et al., eds., *Energy, Productivity, and Economic Growth*, Cambridge: Oelgeschlager, Gunn, and Hain, 1983, 203–14.

______ **and Netschert, Bruce C.**, with V. E. Eliasberg, J. Lerner and H. H. Landsberg, *Energy and the American Economy, 1850–1975*, Baltimore: Johns Hopkins University Press, 1960.

The Response of Energy Demand to Higher Prices: What Have We Learned?

By JAMES L. SWEENEY*

In the decades prior to 1973, demand for energy, particularly energy liquids and electricity, expanded exponentially. From 1950 to 1973, U.S. aggregate energy use grew by 3.5 percent per year, roughly matching real *GNP* growth (3.7 percent). Demand shifted toward petroleum (4.3 percent annual growth) and electricity (7.7 percent) and away from coal (1.0 percent annual growth). These trends, reflected throughout the world, were encouraged by flat or gradually declining energy prices.

Since 1973 energy price increases have been pervasive, although increases have differed radically among energy carriers and over time, being largest since 1979. From 1973 to 1982, real gasoline prices to consumers have increased 51 percent, natural gas delivered to households 139 percent, and residential electricity average price only 23 percent. Fuel prices delivered to electric utilities have varied more: oil increased 175 percent, natural gas, 350 percent, while coal price increased but 85 percent.

Energy demand adjustments have been profound. Between 1973 and 1982, U.S. consumption of oil and gas has declined, oil averaging a 1.4 percent annual decline, and natural gas, 2.3 percent. Electricity growth has been reduced to 2.1 percent per year. Of the fossil fuels, only coal demand has been encouraged, rising at an average annual rate of 2.6 percent. The U.S. primary energy use declined at an annual rate of 0.6 percent. And measured after subtracting electric utility conversion losses (secondary energy), use declined 1.3 percent annually since 1973. Noncommunist world oil consumption declined from 48 million barrels per day (mmb/d) in 1973 to 46 mmb/d in 1982, and primary energy consumption has fallen annually since 1979. On the other hand, coal consumption increased an average of 2.3 percent per year. While demand reductions have been unsteady, occurring primarily since 1979, the old growth patterns have clearly been radically altered.

Thus following large increases in energy prices, demands have shifted toward coal, the fuel experiencing the smallest price increase, and away from oil and gas, fuels experiencing the greatest increases. Overall energy demand growth has been severely curtailed.

A natural experiment was created allowing examination of sensitivity of energy consumption to prices. However, there have been government- and utility-sponsored conservation programs, information campaigns, tax incentives, and moral suasion. Since 1973, U.S. economic growth has diminished, with average real *GNP* growth only 1.8 percent per year. We are just recovering from a recession involving low capacity utilization. Thus although we have accumulated a vast body of evidence, that evidence is subject to a variety of interpretations.

My goal is to communicate key conclusions learned partially through the natural experiment and partially through information available prior to 1973. The exposition progresses from very general conclusions about the nature of energy demand toward more specific quantification of energy demand elasticities.

I. General Conclusions

Aggregate statistics suggest there have been significant reductions in energy use, referred to as energy conservation, and significant

*Energy Modeling Forum, Terman 406, Stanford University, Stanford California 94305.

shifts from one energy carrier to another, referred to as interfuel substitution.

PROPOSITION 1: *Demand responses to higher energy prices — energy conservation and interfuel substitution — typically involve substitutions of other factors for energy, one energy carrier for another, or both mechanisms.*

Study of demand responses involves study of substitution mechanisms—an idea subtly different from assumptions that energy conservation is reduction of "waste" in use of energy or "fat" in the system. Most involve substitution of one input for another: capital for energy, labor for energy, one form of energy for another, or, perhaps most commonly, combinations of capital, labor, and energy for other combinations having different factor proportions.

To understand substitution processes, one should understand several fundamental characteristics of energy use. The first can be summarized:

PROPOSITION 2: *In virtually no uses are energy commodities desired for themselves. Energy demand is derived from demand for more basic end products: for example, warm or cold space, process heat, transportation, light, or motive power.*

Many substitute processes to produce the same end product could use less energy or could use different energy carriers. Energy price increases motivate firms and consumers to select such substitute processes and thereby conserve energy or substitute fuels, even without changing the mix of end products. For example, insulation allows the same amount of warmth to be produced using less energy, more capital, and more labor to install the insulation. Natural gas, oil, or electricity can heat interior space.

Substitution among end products also occurs. When energy prices rise, end products characterized by high energy factor proportions (either used directly or indirectly) increase in price, and users substitute from these products, reducing energy demand and demand for complementary products. For example, when electricity prices increase, cost of air-conditioned space rises, firms and consumers use less, reducing demand for appliances and electricity.

PROPOSITION 3: *Energy is used in virtually all economic activity, but factor proportions vary widely. Thus use of nonenergy commodities involves indirect energy use, but quantities of such embodied energy vary greatly.*

Energy demand may be reduced even without change in energy used directly per unit of any activity. Increasing energy prices increase relative price for commodities embodying higher than average energy intensity and motivate substitutions from these commodities, thereby reducing energy demand in the economy. For example, transportation costs rise relative to telecommunication costs as energy prices rise, firms substitute telecommunication for transportation, and overall energy demand in the economy decreases.

This proposition implies that the energy demand elasticity for the entire industrial sector or economy can be greater than the energy demand elasticities averaged over all individual industries.

PROPOSITION 4: *Most energy is used in conjunction with long-lived capital equipment which, once in place, has fairly fixed energy requirements per unit of equipment use.*

Mechanisms underlying short-run responses may be qualitatively different from those underlying long-run responses. Short-run responses are primarily related to changing intensity of utilization of energy using equipment and only secondarily related to changing factor proportions. In the long run, not only can usage intensity be changed, but also factor proportions of the new energy using capital equipment may be quite different from those of the old equipment it replaces.

Equipment quantities can vary radically. For example, existing automobiles can be driven less, and fuel efficient cars can be driven more, relative to less efficient ones. Maintenance can influence fuel performance, but fuel efficiency of existing autos does not change greatly in response to energy price increases. In the longer run, fuel efficiency of new cars may be fundamentally altered when energy prices rise.

Although this proposition and conclusions have been empirically supported for gasoline in automobiles, electricity use by household appliances, residential space heating, and electricity generation, it has not been fully tested. But I expect further generalization.

PROPOSITION 5: *Long-run energy price adjustments tend to be substantially greater than adjustments occurring over several years or even a decade. Thus conservation and interfuel substitution motivated by price rises can continue to increase for many years after prices stop rising.*

This result, suggested by differences in long- and short-run adjustment processes, follows when the capital stock turns over slowly and when capital stock changes can quantitatively dominate utilization changes.

Gasoline use by automobiles is a good example. Fuel efficiency of the average automobile in 1973 was 13 miles per gallon (mpg). Although new car on-the-road performance increased to 23 mpg by 1982, the average automobile in 1982 obtained only 16 mpg, since the automobile stock is dominated by older cars. In long-run equilibrium, if new car efficiency remains at the current level, average efficiency will increase toward 23 mpg. Thus long-run demand response through auto stock efficiency would more than double the response to date. Even if gasoline prices fall and new car efficiency drops somewhat, gasoline consumption per mile of driving would continue to decline.

The same phenomenon occurs for other uses. Survey data suggest that thermal properties of new U.S. homes heated by natural gas imply an average annual average consumption of 190 mmBtu/1000 square feet, compared to 280 in 1973. Since new homes constructed annually represent only 1–2 percent of the stock, only a small proportion of the ultimate adjustment has been seen to date.

Limits on capital equipment supply can further reduce adjustment speed. For example, short-run limits on the rate of insulation manufacture and installation reduce capital stock adjustment speed and further lengthen adjustment time.

Development of new technology responds to higher energy prices. This process involves very long lags: research and development, testing, commercialization, capital stock turnover, and finally energy use. High energy prices have launched such processes and adjustments will continue for years. Examples include electronic monitoring, control, and software for HVAC systems, improvements in internal combustion engines, more efficient electric motors, solar panels for homes, and newly designed fuel efficient airplanes.

PROPOSITION 6: *Although adjustment of demand to higher prices can be expected to be slow, the precise adjustment rates are unknown.*

Purely econometric studies using any common distributed lag functional forms are weak for estimating dynamics. Process models, theory, and end use data suggest slow adjustment, but all are based upon assumptions about short-run changes in utilization intensities and factor proportions.

Adjustment speed conclusions are important for understanding future effects of energy price changes and for interpreting recent history. The higher the adjustment speed, the greater the fraction of long-run adjustments experienced so far, the smaller the adjustments still to be expected, and the lower the implied long-run demand elasticity.

Other dynamic issues may be important. Economic downturns characterized by large drops in capacity utilization may lead to temporary energy demand reductions which will disappear when capacity utilization again increases.

Because post-1973 energy-using equipment generally uses less energy per unit of output than equipment from older vintages, input factor proportions vary widely in the existing capital stock. This capital heterogeneity implies that declines in capacity utilization allow firms to select *which* units to temporarily retire. Most likely to be withdrawn is older equipment with a high ratio of energy use to output. Thus declines in capacity utilization will reduce energy use standardized for output level; subsequent utilization increases should reverse the reduction. I speculate that the large drops in capacity utilization in manufacturing and electricity generating sectors during the recession have in this way temporarily reduced energy demand.

Because faster economic growth implies faster turnover of the capital stock, whenever new capital stock is more efficient than old, faster growth in economic activity implies more rapid adjustment and therefore declines in energy use per unit of economic activity. This phenomenon helps to explain Japan's impressively rapid adjustment to higher energy prices.

PROPOSITION 7: *Energy consumption occurs at some location. Changes in economic activity may change the location of energy consumption but may or may not influence total energy consumption.*

Migration of industries between nations or regions may occur in response to energy price differences. Such migration reduces energy consumption in the initial location, but may increase, decrease, or leave unchanged energy consumption overall. Such migration will occur in response to energy price locational differentials but not to general increases in energy prices.

This proposition has methodological significance. Empirical cross-section studies measure long-run demand differences motivated by locationally differentiated prices. These studies thus implicitly include interregional migration motivated by price differentials. Because a general increase in energy prices will not create such differentials, such studies may overstate long-run price elasticities relevant to a general increase in energy prices.

I now turn to quantification of elasticities. Aggregate responses will be discussed prior to fuel specific responses. Except where noted, all quantitative estimates refer specifically to the United States.

II. Aggregate Responses to Energy Price Increases

The aggregate elasticity of energy demand measures the impact of energy prices on the consumption of all energy. The term "aggregate elasticity" implies a rule or index for aggregating prices and quantities of various energy commodities. The precise rule is important because energy prices and quantities do not all change in the same proportions.

A common quantity aggregator uses heat values per unit of specific energy carriers to convert physical quantities into total quantities of energy. Such a sum is typically expressed in millions of Btu's (mmBtu) or quadrillions of Btu's (quads). Alternatively, one could treat energy like any other commodity and aggregate using price and quantity weights, for example, Paasche, Laspeyres, Ideal, or Tornquist indices. In what follows a Tornquist index is used. Although results would be virtually invariant among these four indices, choice of heat values as an aggregator could change the measured elasticity.

Total energy use can either include conversion losses in electricity generation (generally referred to as primary energy) or exclude these losses (secondary energy). Demand elasticity varies systematically depending upon where prices and quantities are measured. Elasticities measured at the point of first production (primary energy) will be lowest, those measured at the busbar and refinery gate (secondary energy) will be higher, while those measured at the point of delivery (delivered energy) will be the highest.

PROPOSITION 8: *The long-run aggregate elasticity of demand for secondary energy is likely to be in the range of* -0.4 *to* -0.7. *Measured at the point of delivery, the probable range is* 25 *to* 50 *percent higher, while at the primary level the probable range is* 10 *to* 30 *percent lower.*

Since primary and secondary elasticities are probably smaller than unity, increases in energy prices can be expected to increase the share of output used to pay for energy costs even in the long run.

Propositions 5, 6, and 8 all suggest that under current prices, energy demand adjustments are far from over. The producer price index for fuel and power increased relative to the *GNP* implicit price deflator by a factor of 2.6 from 1972 to 1983. The −0.4 to −0.7 range suggests a 32 to 49 percent long-run reduction of secondary energy use per dollar of *GNP*. The U.S. consumption per *GNP* dollar has decreased 25 percent, significantly less than the long-run estimate.

What is the basis of Proposition 8? Secondary demand elasticity estimates were derived through the Energy Modeling Forum study (1981), but the general range is consistent with much evidence. See William Hogan (1979, 1983) for more detailed discussion. Econometric studies based upon U.S. data—either time-series or cross-section—have been conducted using a variety of techniques and give results well within this range. Cross-sectional studies among OECD countries tend to provide the highest estimates. In addition, more detailed international comparisons, particularly between Sweden, the United States, and Canada suggest much room for further North American adjustment. Structural models suggest that factor proportions can adjust greatly in the U.S. economy. Evidence from engineering studies of specific technologies such as refrigerators and detailed studies of energy consuming sectors show that much energy conservation and fuel switching has occurred and is continuing to occur, and that at current prices much more is cost effective. All this evidence is generally consistent with the cited range and strongly supports the proposition that more adjustments are to come.

Factors in Proposition 8 for translating to delivered or primary elasticity are still fairly judgmental. They are based upon estimated cost markups, including taxes, at different stages of the supply chain.

Aggregate elasticity must be used with caution. Aggregate elasticity is sensitive to precise composition of price changes and is only a rough guide.

One complicating factor which could potentially make analysis of the historical record more difficult is the role of government programs:

PROPOSITION 9: *The extent to which government-sponsored energy conservation programs or other nonmarket forces have reduced the demand for energy is unknown. However at least* 80 *percent and probably much more of the demand reductions can be attributed to price and economic activity changes.*

Educated opinions vary on the effects of the federal energy conservation programs and other nonmarket forces such as utility conservation programs and fear of shortages. Econometric estimates attribute to nonmarket forces as much as 20 percent of conservation occurring to date (Eric Hirst et al. 1983), or as little as 0 percent (Hogan, 1983). Department of Energy (1982) estimates based upon program reviews suggest less than 5 percent of the observed conservation is based upon federal programs, but ignore the role of the fuel efficiency standards for new cars (CAFE standards).

Analysis of nonmarket forces is necessary for interpreting the historical record. The greater the role of forces other than prices and economic activity, the smaller the market response that can be inferred from observed energy consumption reductions. However, even using the highest estimates of nonmarket forces, at least 80 percent of the observed demand adjustment has resulted from price and economic growth changes.

PROPOSITION 10: *The extent to which the current recession has contributed to the reduced energy demand is subject to debate.*

As discussed above, with heterogeneous capital stock, low capacity utilization could reduce energy demand. How much of the current industrial sector demand reduction can be attributed to this phenomenon is not clear, although some electric utility fuel shifting undoubtably can be. Thus, to the extent the recession has contributed to the sharp drop in oil and energy use of the last two years, these recent data may overstate the demand reduction to be expected for the next few years, although not for the long run.

Propositions 6, 9, and 10 help explain why one cannot simply infer long-run price elasticities from measuring demand reductions occurring to date. Some reductions are due to factors other than economic activity and price; we cannot be sure how much reduction is due to changes in economic growth or to the recession; we do not know adjustment speeds precisely and thus do not know the ratio of long-run elasticity to short- or intermediate-run elasticities.

III. Fuel Specific Price Responses

Since energy price changes can lead to conservation and interfuel substitution, demand for an energy carrier will be decreasing in its price and increasing in prices of competing fuels: own elasticities will be negative and cross elasticities will be positive. The aggregate elasticity must be smaller than the weighted average of fuel specific own elasticities. Thus the fuel specific secondary elasticities should on average be more negative than the -0.4 to -0.7 range and delivered energy elasticities should on average be more negative than the -0.5 to -1.0 range.

PROPOSITION 11: *The long-run delivered price elasticity of demand for electricity probably exceeds unity but may be as low as* -0.7.

While most studies suggest elasticities exceeding unity, many careful studies estimate lower figures (Douglas Bohi, 1981; Lester Taylor, 1975). But these econometric studies would not be expected to capture effects of new technologies created in response to increases in electricity prices. While the lower estimates cannot be rejected, the higher elasticity is more probable.

If the elasticity does exceed unity, electricity price increases may not increase long-run revenue for utilities; however, such increases can improve the financial situation of regulated utilities by greatly reducing or eliminating the need for new capital equipment.

Consistent with economic theory, marginal electricity prices empirically seem more important than average prices in determining demand. In recent years many public utility commissions have ordered utilities to structure rates to eliminate declining block structures and to introduce "lifeline" and other increasing block rate structures. These changes may have contributed to the demand growth decline and may continue to do so.

The many studies of electricity demand have collectively established that estimated elasticities can vary widely when different data bases or methods are used (see Bohi). Thus demand growth uncertainty remains. Since uncertainty will be greatest for equipment requiring long lead times, such as coal-fired or nuclear generators, I expect utilities to continue to shun such projects.

PROPOSITION 12: *We have only poor information on the long-run demand elasticities for natural gas.*

There has virtually never been an unconstrained market for natural gas; in early years pipelines were being constructed and access was limited, in later years shortages precluded new hookups. Thus econometric studies could capture the role of prices in influencing conservation but underestimate interfuel substitution. Therefore we have only poor information about demand for natural gas and about the effect of natural gas prices on demands for other energy carriers.

PROPOSITION 13: *The demand for oil, natural gas, and coal in the industrial sector (including electricity generation) is a highly nonlinear function of price of those fuels and of competitive fuels.*

Many electric utility or industrial boilers are fitted to use either oil or gas, or sometimes coal slurries. In the longer run, new construction also allows coal to be used for large boilers. Therefore direct interfuel substitution can be extensive when fuel prices are nearly equivalent on a cost per Btu basis. But for prices significantly different from Btu equivalence, almost no interfuel substitution can be expected. Thus own and cross elasticities will be very large at prices which could motivate the substitution and virtually zero elsewhere.

PROPOSITION 14: *There is almost no possibility for interfuel substitution in nonrail transportation*; *thus the demand for petroleum in this sector will be virtually independent of other energy prices.*

Energy conservation can be expected to dominate with little interfuel substitution since liquid fuels are virtually required for these transportation activities.

PROPOSITION 15: *Short-run delivered elasticity of demand for gasoline is near* -0.2 *and long-run elasticity is probably in the* -0.6 *to* -1.0 *range.*

Petroleum is the predominant fuel for transportation, with gasoline for automobiles the largest component. The demand elasticity through automobile utilization is low (in the range of -0.2) while the elasticity through changes in automobile efficiency is far higher (in the range of -0.7). Elasticity estimates are far more consistent across studies than for other fuels.

IV. In Summary

There remains much quantitative uncertainty about responses of energy demand to higher prices and thus much opportunity for further research. However, the natural experiment has made virtually unrefutable the proposition that price changes have motivated much conservation and interfuel substitution, and that much more adjustment should be expected over the years.

REFERENCES

Bohi, Douglas R., *Analyzing Demand Behavior. A Study of Energy Elasticities*, Baltimore: Johns Hopkins University Press, 1981.

Hirst, Eric et al., "Recent Changes in U.S. Energy Consumption: What Happened and Why," Oak Ridge National Laboratory, February 1983.

Hogan, William W., "Dimensions of Energy Demand," in *Selected Studies on Energy: Background Papers for Energy: The Next Twenty Years*, Cambridge: Ballanger, 1979.

______, "Patterns of Energy Use," mimeo., Energy and Environment Policy Center, Harvard University, October 1983.

Taylor, Lester, "The Demand for Electricity: A Survey, *Bell Journal of Economics*, Spring 1975, *6*, 74–110.

EMF 4 Working Group, "Aggregate Elasticity of Energy Demand," *Energy Journal*, No. 2, 1981, *2*, 37–75.

U.S. Department of Energy, Office of Policy, Planning and Analysis, *Sunset Review, Program-by-Program Analysis*, DOE/PE-0040, February 1982.

Supply Shocks and Monetary Policy Revisited

By ROBERT J. GORDON*

A macroeconomic supply "disturbance" or "shock" is any event which creates an autonomous shift in the aggregate supply curve relating the economywide price level to the level of output or utilization. The autonomous nature of such shifts distinguishes them from other movements in the supply curve that represent the consequences of a current or prior changes in aggregate demand. The distinction between supply and demand shocks is valid only with reference to their *origin*, whereas the *consequences* of supply shocks for output and inflation depend fundamentally on the aggregate demand policies that are pursued in their wake.

This paper is written almost a decade after the first attempts in 1974 to develop a theory of policy response to supply shocks.[1] It provides a simple algebraic framework that facilitates a summary of the central issues posed by supply shocks for macroeconomic policy. Primary emphasis is placed on the case for and against monetary accommodation, on the nature and extent of wage indexation, and on the distinction between permanent and transitory shocks. A tight space constraint precludes more than passing mention of cost-oriented fiscal policy, oil tariffs, buffer stocks and other policies that mainly influence the magnitude of the shocks themselves rather than their consequences for macroeconomic performance. Given the difficult tradeoffs faced by monetary policymakers considering the merits of accommodation, these supply-side alternatives may actually represent the best available policy options. The first line of defense against a real disturbance is a real policy.

I. A Simplified Hybrid Model

The original case for the monetary accommodation of an adverse supply shock, as developed by my 1975a paper and by Edmund Phelps, rests on a "macroeconomic externality," that is, a spillover from the unavoidable loss of output in the shocked sector of the economy to a loss of output in the unshocked sector that may be avoidable by monetary accommodation. The case for accommodation is strongest in a model with rigid or sluggishly adjusting nominal wages in the unshocked sector, is weaker in the presence of partial wage indexation, and is nonexistent in the presence of complete wage indexation or instantaneous market clearing achieved by perfectly flexible wages. Here I set out a hybrid model, sharing Phelps' one-sector production technology with my exogenous nominal *GNP* assumption, that allows the analysis of macroeconomic externalities and monetary accommodation to be presented in a more transparent fashion than in the two original papers.

Consider an economy that produces output (Q) using only labor (N) and a raw material (σ):

$$(1) \quad Q = F(N, \sigma), \qquad F_N > 0, F_\sigma > 0.$$

The supply of labor in the economy is fixed at N^*, and so "natural" (or "full employment" or "potential") output is

$$(2) \qquad Q^* = F(N^*, \sigma).$$

Note that no capital is used in production. Capital appears in Phelps' model, but its only role there is to introduce a set of com-

*Northwestern University and National Bureau of Economic Research. Department of Economics, Northwestern University, Evanston, IL 60201. This research has been supported by the National Science Foundation. I am grateful to Robert Flood for helpful discussion, and to Glenn Hubbard and Allan Meltzer for comments on a first draft.

[1] Edmund S. Phelps (1978, p. 206) lists the 1974 conferences at which he and I independently developed what Edward Gramlich later called the "Gordon-Phelps model." I discovered after writing this paper that Stanley Fischer (1983) developed an analysis that is compatible with my Section I but is both more complex and more general.

plex and ambiguous impacts of supply shocks on the real rate of interest and on velocity. Here these second-order effects are neglected through the assumption that nominal *GNP* (Y) is exogenous. The economy's demand price (P^d) is then simply nominal *GNP* divided by actual real *GNP*:

$$(3)\qquad P^d = YQ^{-1} = Y[F(N,\sigma)]^{-1}.$$

Assuming that the product market always clears and labor is paid its marginal product, the economy's supply price (P^s) is equal to the nominal wage rate divided by the marginal product of labor:

$$(4)\qquad P^s = W[F_N(N,\sigma)]^{-1}.$$

The conditions for a macroeconomic externality can now be examined by subjecting this economy to a single comparative static experiment, a change in the raw material input σ, caused by some unexplained event. A macroeconomic externality is defined as occurring when, starting in equilibrium with $Q = Q^*$, the percentage change in Q needed to keep $P^d = P^s$ is not equal to the change in Q^*. Here I shall use the "dot" notation for percentage changes ($\dot Q = dQ/Q$), and so the difference between the rate of actual and natural output change is, from (3):

$$(5)\qquad \dot Q - \dot Q^* = \dot Y - \dot P^d - \dot Q^*.$$

The condition necessary for this to be zero can be worked out by setting $\dot P^d = \dot P^s$, and by noting that if the change in actual *GNP* is equal to that in natural real *GNP*, then both output change terms can be evaluated by assuming that labor input remains at N^*, that is, that $\dot N = 0$. We have from (2) and (4):

$$\begin{aligned}(6)\qquad \dot Q - \dot Q^* &= \dot Y - \dot P^s - \dot Q^* \\ &= \dot Y - \dot W + (F_{N\sigma}/F_N)\,d\sigma \\ &\quad - (F_\sigma/F)\,d\sigma.\end{aligned}$$

Thus the condition for real *GNP* to remain at equilibrium can be written

$$(7)\qquad \dot Y - \dot W = -(F_{N\sigma}/F_N - F_\sigma/F)\,d\sigma.$$

That is, that the *difference* between the percentage change in nominal *GNP* and that in the nominal wage rate remain equal to the right-hand side of (7).

And what is this unfamiliar-looking term? We can write the income share of the raw material (α) as unity minus the share of labor: $\alpha = 1 - F_N N/F$, so that $\dot\alpha = -(\dot F_N + \dot N - \dot F)$. Because at Q^* there is no change in labor input ($\dot N = 0$), the change in the raw material share is just

$$\begin{aligned}(8)\qquad \dot\alpha &= -(\dot F_N - \dot F) \\ &= -(F_{N\sigma}/F_N - F_\sigma/F)\,d\sigma.\end{aligned}$$

Thus substituting (8) into (7), we have the condition:

$$(9)\qquad \dot Y - \dot W = \dot\alpha.$$

While it is completely consistent with the analysis in the original Gordon and Phelps papers, the appeal of (9) is that it is both simpler and more general. There is no need to assume that nominal *GNP* or the nominal wage rate is fixed. Condition (9) applies to either a market-clearing or non-clearing economy. In a market-clearing economy the perfectly flexible wage can adjust downward by any amount needed to open up the required "wedge" between dY/Y and dW/W when the raw material share increases, and there is no necessity for monetary accommodation. However, a rigid or sticky nominal wage rate and an increase in the raw material share together imply that full employment can be maintained only if policymakers generate a sufficient increase in nominal *GNP*. And the relevance of an increasing share is clear, given the actual tripling of energy's value share between 1972 and 1981.[2]

II. Accommodation and Indexation

The theory of monetary policy responses to supply shocks is clear-cut in unrealistic

[2] The share index is calculated by multiplying total real energy consumption by the composite energy deflator (both from the *Statistical Abstract of the United States*, 1982–83, pp. 572–73), dividing by nominal *GNP*, and setting 1972 as the base of the index.

extreme cases and ambiguous in more realistic intermediate cases. Here I ignore effects of supply shocks on the velocity of money, allowing us to link central bank control of the money supply with control over the growth rate of nominal *GNP* ($\dot{Y}_t$). Effects of indexation are examined in a mechanical adjustment equation which allows changes in wage rates to depend only on current and past price changes, on past wage changes, and on the output ratio (Q_t/Q_t^*):

$$(10)\quad \dot{W}_t = \beta\dot{P}_t + \gamma\dot{P}_{t-1} + (1-\beta-\gamma)\dot{W}_{t-1} + \phi(\dot{Q}_t/Q_t^*).$$

This equation is not intended to represent the outcome of maximizing behavior, but rather to allow examination of a taxonomy of consequences of an accommodating monetary policy that maintains full employment, that is, $Q_t = Q_t^*$. In each of the following cases, I normalize on an assumed situation in the period prior to the shock in which $\dot{W}_0 = \dot{Y}_0 = \dot{Q}_0^* = 0$, and assume that the supply shock has a permanent impact on the level of the raw material share only in period one ($\alpha_0 < \alpha_1 = \alpha_2 = \ldots = \alpha_n$). Thus the only nonzero value of $\dot{\alpha}_t$ is $\dot{\alpha}_1 > 0$. Note also that for full employment to be maintained, $\dot{P}_t = \dot{Y}_t - \dot{Q}_t^*$. Substituting (10) into (9) gives

$$(11)\quad \dot{Y}_t = \big[\dot{\alpha}_t - \beta\dot{Q}_t^* + \gamma(\dot{Y}_{t-1} - \dot{Q}_{t-1}^*) + (1-\gamma-\beta)\dot{W}_{t-1} + \phi(Q_t/Q_t^*)\big]/1-\beta.$$

When wage changes depend only on their own past values and on the output ratio ($\beta = \gamma = 0$), full monetary accommodation is clearly optimal. During period one $\dot{W}_1 = 0$, so that an accommodative policy would set $\dot{Y}_1$ to equal $\dot{\alpha}_1$. The opposite extreme occurs with complete indexation of wage changes to current changes in the price level, $\beta = 1$ while $\gamma = 0$. Now the right-hand side of (11) becomes infinite, implying that there is no change in nominal *GNP* that will maintain full employment. Full indexation in the presence of supply shocks is clearly suboptimal, as pointed out by Joanna Gray (1976) and by Stanley Fischer (1977).

Another possible case is that wage changes are indexed fully to lagged price change ($\beta = 0$ while $\gamma = 1$). In this case (11) reduces to the following, when we note that from (2) that $\dot{Q}_t^* = \dot{F}_t$:

$$\dot{Y}_t = \dot{Y}_{t-1} + \dot{\alpha}_t - \dot{F}_{t-1}.$$

In the example of a one-period supply shock, in the first period, $\dot{W}_1 = 0$, and this requires the same accommodative policy as if $\gamma = 0$, that is, $\dot{Y}_1 = \dot{\alpha}_1$. In the second period, however, lagged indexation prevents nominal wage and *GNP* growth from returning to zero. Instead, from (8),

$$\dot{Y}_2 = \dot{W}_2 = \dot{Y}_1 - \dot{F}_1 = \dot{\alpha}_1 - \dot{F}_1 = -\dot{F}_{N_1}.$$

In all future periods,

$$\dot{Y}_t = \dot{W}_t = \dot{P}_{t-1} = -\dot{F}_{N_1}.$$

That is, maintenance of full employment requires a permanent acceleration of inflation and in the growth of nominal wages and *GNP* following any supply shock that permanently shifts the raw material share. In this plausible case of lagged indexation, supply shocks pose a tradeoff between a permanent acceleration of inflation and a temporary loss of output. The severity and duration of the output loss depend on the Phillips curve parameter ϕ or, more generally, on the economy's "sacrifice ratio" (my article with Stephen King, 1982). For the U.S. case I showed (1982, p. 134) that an accommodative policy that cumulatively raised the money supply by 9 percent in 1975–80 compared to an alternative hypothetical constant-growth money path would have resulted in 1.9 percentage points more inflation in 1980 with the benefit of 3.2 fewer point-years of unemployment during 1975–80 (an output gain of 8 percent of a year's *GNP*).

In the realistic case of a permanent shock and partial and/or lagged wage indexation, the optimal degree of accommodation depends on a finely balanced comparison of the welfare costs of inflation and unemployment. The optimal outcome is different in a society like the United States in 1973–75, where inflation had high costs due to nonneutral tax rules and binding financial rate ceilings, than in a society like Israel or Brazil, in which real interest rates and tax rates were much more neutral with respect to inflation. In a sense there is a cumulative interaction, as I suggested earlier (1975b), between monetary accommodation, behavior regarding contract lengths and the Phillips curve parameter (ϕ above), and institutional rules regarding tax rates and financial regulations. Inflation begets a neutralized institutional environment, which begets accommodation and more inflation.

III. The Persistence of Shocks and the Formation of Expectations

In the above example an adverse supply shock causes a permanent reduction in the economy's productive capacity. Another possibility is that the shock is temporary, as in the case of an agricultural drought or freeze. In this case the tradeoff with partial or lagged indexation is between a temporary output loss and a temporary rather than permanent acceleration of inflation. Even a temporary upsurge in the inflation rate is not without welfare costs, since it causes a permanent increase in the price level at every date in the future and a corresponding loss in the wealth of holders of high-powered money (effects on interest-bearing assets and liabilities cancel out).

Thus far nothing has been said about inflation expectations. If the indexation parameters β and γ are set by legislation, then wage changes would evolve mechanically in the aftermath of a supply shock, as described above. If β and γ are relatively low at the time of the shock, for example, if wage changes are determined mainly by their own past values, then the decline in the real wage rate associated with the shock may create political pressure to have indexation legislation changed. Indeed the percentage "pass through" of price changes in the Italian *scala mobile* indexation agreement was raised in 1975 after the first oil shock. However, in most countries indexation parameters are not set in legislative stone, but are subject to frequent negotiation between workers and firms. Multiperiod wage agreements achieved in delicate negotiations would not tend to be altered in response to a temporary shock that is expected to leave output and the real wage unaffected after a transition period of a few months or a year.

But a shock expected to have a permanent effect on output and the real wage poses a serious dilemma for the parties in wage negotiations, and may well lead to a change in any or all of the parameters of (11). As depicted in the model of John Taylor (1980), newly negotiated contracts depend not just on the current state of demand, as in (11), but also on the expected *future* state of demand. Taylor's agents are "forward looking," not "backward looking" as in mechanical formulae like (11). Workers with forward-looking expectations can calculate the future consequences of maintaining high β and γ indexation parameters in the face of a permanent supply shock—permanently higher inflation if the policy authorities accommodate, and a period of low aggregate demand (Q/Q^*) if they do not accommodate. Faced with this unpleasant tradeoff, rational workers would suspend indexation and allow the real wage to fall by the required amount. Hence the rational expectations response to a permanent shock merges together with the market-clearing outcome described above.

The painless transition implied by quickly adjusting forward-looking expectations to a permanent shock has not been observed in fact. As Jeffrey Sachs (1979) has emphasized, unemployment increased in virtually all OECD countries after the 1973–74 oil shock, reflecting a combination of nonaccommodative aggregate demand policies, and an excess of real wage growth over productivity growth. One possible explanation for this outcome is that economic agents initially thought the oil shock would be temporary and were slow to learn that it was permanent. Karl Brunner, Alex Cukierman, and Allan Meltzer (1980) show that, even within the context of a market-clearing model, a permanent reduction in productivity can cause stagflation, because agents only gradually learn the permanent values of real variables and only gradually adjust their anticipations. Consistent with their analysis is my 1983b finding that real wage growth in most large European countries was much more moderate after the 1979–80 oil shock than after the initial 1973–74 shock. Having seen the effects of the first shock persist, agents were more prepared to believe that the second would persist as well.

IV. Impact on Doctrinal Debates

Supply shocks have helped to unify the teaching of macroeconomic theory with that of microeconomics, since basic results in both subjects can be summarized with supply and demand curves. Undergraduates are now taught that unemployment and inflation may be either negatively or positively correlated. Following an autonomous shift in demand, the extent and duration of any change in unemployment depends on the length of wage contracts and the adjustment of expectations, while following an autonomous shift in supply, the extent and duration of any change in unemployment depends on the interaction of wage indexation and monetary accommodation. The recognition that inflation depends on shifts in both demand and supply, not just on past changes in the money supply, has facilitated econometric explanations of the inflation process that appear able to explain why in the 1970's U.S. inflation was so variable and why in 1981–83 it decelerated so rapidly.[3]

The positive correlation of inflation and unemployment in the 1970's brought forth many responses. In a famous polemic (1978), Robert Lucas and Thomas Sargent used this positive correlation to challenge the application of "Keynesian" models to macroeconomic policymaking. Their stated intent was "to establish that the difficulties are *fatal*: that modern macroeconomic models are of *no* value in guiding policy and that this condition will not be remedied by modifications along any line which is currently being pursued" (p. 50). Especially with respect to the issue at hand, this dismissal is inappropriate. Observations in the inflation-unemployment quadrant can represent the interaction of demand and supply curves. The Lucas-Sargent challenge failed to notice the concurrent development of new "Phillips curve" formulations which combined the effects of supply and demand shifts with that of sluggish price adjustment, the basic element in Keynesian economics. As put forth in my article with King, the U.S. Phillips curve appears to be one of the most stable empirical macroeconomic relationships of the postwar era, one that shows no sign as of yet of being subject to Lucas' econometric critique.[4] In basing their attack on Keynesian economics on the alleged collapse of the Phillips curve, Lucas and Sargent seem in retrospect like teenage pranksters who scare everyone by crying "wolf" and then flee the scene when it is discovered that there is no wolf.

Finally, supply shocks have raised the perennial question of the optimality of decentralized and uncoordinated wage and price setting. Decentralization ("the invisible hand") is usually supported by economists as required for *microeconomic* efficiency, yet coordination and centralization may be needed to obtain an improved *macroeconomic* response to supply shocks. In the past decade economists have debated the merits of alternative responses that would have required coordinated action, including a one-time real wage reduction to match the decline in productivity caused by the 1973–74 and 1979–80 oil shocks, changing indexation formulae to exclude oil prices and indirect taxes from the price measure used for escalation, and oil import taxes balanced by reductions in other indirect taxes to put downward pressure on the world oil price and to discourage consumption.

[3] Models that combine demand and supply elements include those of Otto Eckstein (1980), my 1982 article, and my article with King. Readable descriptions of the role of supply shocks in the inflation of the 1970's are provided by Alan Blinder (1979, 1982). An evaluation of the 1981–83 disinflation is provided by the three papers in the volume edited by William Nordhaus (1983).

[4] The stability of the inflation equation to changes in sample period is examined by myself and King (p. 218) and related to the Lucas critique (pp. 224–29). Structural shifts in the twentieth century prior to 1954 are discussed in my 1983a article and by Meltzer (1977).

REFERENCES

Blinder, Alan S., *Economic Policy and the Great Stagflation*, New York: Academic Press, 1979.

______, "The Anatomy of Double-Digit Inflation in the 1970s," in R. E. Hall, ed., *Inflation: Causes and Consequences*, Chicago: University of Chicago Press (NBER), 1982, 261–82.

Brunner, Karl, Cukierman, Alex and Meltzer, Allan H., "Stagflation, Persistent Unemployment and the Permanence of Economic Shocks," *Journal of Monetary Economics*, October 1980, *6*, 467–92.

Eckstein, Otto, *Core Inflation*, Englewood Cliffs: Prentice-Hall, 1980.

Fischer, Stanley, "Wage Indexation and Macroeconomic Stability," in K. Brunner and A. Meltzer, eds., *Stabilization of the Domestic and International Economy*, Vol. 5, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl. 1977, 107–47.

______, "Supply Shocks, Wage Stickiness, and Accommodation," Working Paper No. 1119, National Bureau of Economic Research, May 1983.

Gordon, Robert J., (1975a) "Alternative Responses of Policy to External Supply Shocks," *Brookings Papers on Economic Activity*, 1:1975, 183–206.

______, (1975b) "The Demand for and Supply of Inflation," *Journal of Law and Economics*, December 1975, *18*, 807–36.

______, "Inflation, Flexible Exchange Rates, and the Natural Rate of Unemployment," in Martin N. Baily, ed., *Workers, Jobs, and Inflation*, Washington: The Brookings Institution, 1982, 89–158.

______, (1983a) "A Century of Evidence of Wage and Price Stickiness in the United States, the United Kingdom and Japan," in J. Tobin, ed., *Macroeconomics, Prices, and Quantities*, Washington: The Brookings Institution, 1983, 85–121.

______, (1983b) "The Wage and Price Adjustment Process in Eight Large Industrialized Countries," Working Paper, October 1983.

______ **and King, Stephen R.,** "The Output Cost of Disinflation in Traditional and Vector Autoregressive Models," *Brookings Papers in Economics*, 2:1982, 205–42.

Gray, Joanna, "Wage Indexation: A Macroeconomic Approach," *Journal of Monetary Economics*, April 1976, *2*, 221–35.

Lucas, Robert E. Jr., and Sargent, Thomas J., "After Keynesian Macroeconomics," *After the Phillips Curve*, Federal Reserve Bank of Boston Conference Series, No. 19, 1978, 49–72.

Meltzer, Allan H., "Anticipated Inflation and Unanticipated Price Change," *Journal of Money, Credit and Banking*, Part 2, February 1977, *9*, 182–205.

Nordhaus, William D., *Inflation Prospects and Remedies*, *Alternatives for the 1980s*, No. 10, Washington, Center for National Policy, 1983.

Phelps, Edmund S., "Commodity-Supply Shock and Full-Employment Monetary Policy," *Journal of Money, Credit, and Banking*, May 1978, *10*, 206–21.

Sachs, Jeffrey D., "Wages, Profits, and Macroeconomic Adjustment: A Comparative Study," *Brookings Papers on Economic Activity*, 2:1979, 269–319.

Taylor, John B., "Aggregate Dynamics and Staggered Contracts," *Journal of Political Economy*, February 1980, *88*, 1–23.

U.S. Bureau of the Census, *Statistical Abstract of the United States*, Washington, 1982–83.

A Simple Account of the Behavior of Long-Term Interest Rates

By JOHN Y. CAMPBELL AND ROBERT J. SHILLER*

To a first approximation, long-term interest rates behave like short-term interest rates. For example, the yields on twenty-year Treasury bonds and on one-month Treasury bills tend to peak and to bottom out together. Thus people often speak of "the level of interest rates" without specifying maturity.

The spread between long and short rates tends to be unusually small or even negative when short rates are high relative to the experience of the last few years. Franco Modigliani and Richard Sutch (1967) showed that the relation between long and short rates can be well described by expressing the long rate as a five-year distributed lag of short rates, with the coefficients summing to about one and with substantial weight on the current short rate. Recent experience upholds this characterization except that the distributed lag has become shorter (Albert Ando and Arthur Kennickell, 1983). Equivalently, the spread between long and short rates is well explained by current and lagged short rates, with approximately equal and opposite coefficients on the current rate and the sum of lagged rates.

This moving average relation could be consistent with the simple expectations theory of the term structure, if investors look to the recent past to form expectations about future interest rates. Whether such expectations are rational depends on the time-series properties of short-term interest rates. Depending on the policy regime and its implications for the movements of short rates, the observed distributed lag might correspond to a rational expectations theory of the term structure, or a theory of overreaction or underreaction of long rates to short rates, relative to the predictions of the rational expectations model. Experimental psychologists, such as Amos Tversky and Daniel Kahneman (1974), claim to have shown that people tend to overreact in their expectations to evidence which seems superficially to be relevant, even after experience should have convinced them otherwise. This suggests that there might be policy regimes where the long rate overreacts to temporary movements in short rates. Of course, any such "overreaction" might also be reconciled with the theory of finance if certain covariances change with the short rate.

A look at the data suggests an abrupt policy shift starting with the Fed's new operating procedures in October 1979. We concentrate here on the policy regime which prevailed between the 1951 Treasury accord and 1979. Modigliani and Shiller (1973) claimed that, for the early part of the period, the observed distributed lag was approximately consistent with the time-series properties of the short rate given a simple expectations model, and Thomas Sargent (1979) was unable to reject this hypothesis with a likelihood ratio test in a vector autoregression. However, more recent work has cast doubt on the notion that the simple rational expectations model of the term structure is adequate even as a first approximation to the behavior of interest rates. It was shown by Shiller (1979) that when long-term interest rates are unusually high relative to short rates, they then tend to fall rather than rise as predicted by the expectations theory. Our 1983 article with Kermit Schoenholtz showed that when six-month bill rates are higher

*Yale University, New Haven, CT 06520. Research support was provided by the National Science Foundation and the Sloan Foundation. A version of this paper with a data appendix is available as National Bureau of Economic Research Working Paper No. 1203.

than three-month bill rates, there is no tendency for the three-month bill rate to rise subsequently. Lars Hansen and Sargent (1981) were able to reject the rational expectations theory at the 0.5 percent level with a likelihood ratio test on postwar U.S. data when an additional restriction involving the current long-term interest rate was added to Sargent's earlier formulation.

These results might be summarized as finding that the behavior of long-term interest rates is dominated by a "risk premium" which is so variable as to swamp out expectations in determining the slope of the term structure. The phrase risk premium has been defined in various ways in the term structure literature. We turn next to a discussion which will clarify the relations among these definitions. This enables us to state more formally the hypotheses that long rates overreact or underreact to short rates, and it provides a framework in which we characterize interest rate behavior.

I. "Well-Tempered" Definitions of Risk Premia

We make use here of approximations to holding-period yields and forward rates which are obtained by linearizing the exact expressions around the coupon rate on a long-term bond. These approximations were developed by ourselves and Schoenholtz. We also investigated their accuracy. Without such preliminary linearization, small differences among alternative definitions of risk premia, arising from nonlinear functions in expectations, make it difficult to consider the definitions within a single framework. The analogy with the approximation which allows a musical instrument to be tuned to more than one key at a time leads us to call our system a well-tempered one.

We chose our definitions to facilitate comparison with bond yields as commonly quoted. Bonds issued with less than a year to maturity commonly carry no coupons, but longer-term bonds generally pay coupons which bring their sale price near par. It is natural then to define the five-year ahead, ten-year forward rate, for example, as the yield on a ten-year coupon bond to be purchased at par five years hence. Such an asset can be constructed today as a portfolio of bonds with maturities up to fifteen years. Similarly, the five-year holding return on a fifteen-year bond is the yield to maturity on buying the fifteen-year bond, receiving its coupons, and selling it five years hence (when it is a ten-year bond). The fifteen-year holding yield on a five-year bond is the yield to maturity on an investment in three consecutive five-year coupon bonds, reinvesting principal (i.e., rolling over the five-year bonds) but receiving coupons.

The linear approximation to the j-period holding yield on an i-period bond is

$$(1)\quad h_t^{(i,j)} = \left(D_i R_t^{(i)} - \left(D_i - D_j\right) R_{t+j}^{(i-j)}\right)/D_j, \qquad 0 < j \leq i$$

$$(2)\quad h_t'^{(i,j)} = \left(1/D_j\right)\left[\sum_{k=0}^{(j-i)/i} \left(D_{ki+i} - D_{ki}\right) R_{t+ki}^{(i)}\right] \qquad 0 < i \leq j,\ j/i \text{ integer}$$

The linear approximation to the n-period ahead m-period forward rate is

$$(3)\quad f_t^{(n,m)} = \left(D_{m+n} R_t^{(m+n)} - D_n R_t^{(n)}\right) / \left(D_{m+n} - D_n\right), \quad 0 < m, 0 \leq n$$

where $R_t^{(i)}$ = yield to maturity on an i-period bond; $D_i = (1-g^i)/(1-g)$, $g = 1/(1+\bar{R})$, $\bar{R}$ = coupon rate. D_i is the "duration" of an i-period bond selling at par with coupon $\bar{R}$, as defined originally by Frederick Macaulay (1938). Duration is intended as a better measure than maturity of how "long" a bond is. It takes account of the fact that bonds with coupons derive much of their value from payments which are made earlier than maturity. Thus for bonds with coupons, $D_0 = 0$, $D_{i+1} - D_i = g^i$, so $D_i < i$ for $i > 1$. For pure discount bonds, $\bar{R} = 0$ and duration and time to maturity are the same.

The simple expectations theory of the term structure, with no allowance for risk, equates

$E_t h_t^{(i,j)}$ or $E_t h_t'^{(i,j)}$ with $R_t^{(j)}$, and $f_t^{(n,m)}$ with $E_t R_{t+n}^{(m)}$. Risk premia are deviations from this theory, which can be written either as differences between expected holding returns and yields, or as differences between forward rates and expected spot rates. We denote the former as $\phi^{(i,j)}(j \leq i)$ or $\phi'^{(i,j)}(j \geq i)$, and the latter as $\psi^{(n,m)}$. Then we have the holding-period risk premium:

$$(4) \qquad \phi_t^{(i,j)} = E_t h_t^{(i,j)} - R_t^{(j)}, j \leq i$$

the rolling risk premium:

$$(5) \qquad \phi_t'^{(i,j)} = R_t^{(j)} - E_t h_t'^{(i,j)}, j \geq i$$

and the forward rate risk premium:

$$(6) \qquad \psi_t^{(n,m)} = f_t^{(n,m)} - E_t R_{t+n}^{(m)}.$$

ϕ, ϕ' and ψ all appear in the existing literature on the term structure. Our well-tempered formulation allows us to derive simple linear relationships among them. First, we can substitute (1) and (3) into (4) and (6) to show that

$$(7) \qquad \phi_t^{(i,j)} = (D_i - D_j)\psi_t^{(j,i-j)}/D_j.$$

Second, we can rearrange equation (3) so that it expresses the j-period bond rate as a weighted average of forward rates of maturity i, with weights equal to those in equation (2). It is immediate that

$$(8)$$
$$\phi_t'^{(i,j)} = (1/D_j)\left[\sum_{k=0}^{(j-i)/i} (D_{ki+i} - D_{ki})\psi_t^{(ki,i)}\right],$$
$$0 < i \leq j$$

Finally, we can rearrange equation (1) so that it expresses the j-period bond rate at time t as a function of the i-period holding return on a j-period bond and the $(j-i)$-period bond rate at time $t+i$. By recursive substitution, we obtain the following expression:

$$(9) \quad \phi_t'^{(i,j)} = (1/D_j)\left[\sum_{k=0}^{(j-i)/i} (D_{ki+i} - D_{ki}) \times E_t \phi_{t+ki}^{(j-ki,i)}\right], 0 < i \leq j.$$

A natural interpretation of the notion that long rates overreact to short rates is that long bonds are a "good investment" when the short rate is high. In other words, the returns on long bonds over some holding period tend to be higher than those predicted by the expectations theory when the short rate is high: the holding period or rolling risk premium is positively correlated with the short rate.[1] In the next section we examine the relation between the one-month excess holding return on a twenty-year bond, and the one-month Treasury bill rate. We do not calculate the twenty-year excess return on a twenty-year bond, which includes the rolling risk premium, since we have only just over twenty years of data. However, we study the rolling risk premium indirectly by conducting an ARIMA analysis of the one-month bill rate.

II. The Behavior of Risk Premia

We can estimate ϕ_t by regressing the excess return $h_t^{(i,j)} - R_t^{(j)}$ on variables in the information set at t. The excess return is just $(D_i/D_j - 1)$ times the forward-spot rate difference $(f_t^{(j,i-j)} - R_{t+j}^{(i-j)})$, so equivalent results are obtained with this dependent variable.

[1] N. Gregory Mankiw and Lawrence Summers (1983) interpreted overreaction as the hypothesis that the long rate behaves according to the expectations model for a bond of shorter duration. This definition is consistent with ours, in that if long rates overreact in Mankiw and Summers' sense, and if the time-series process for short rates is stationary, then the holding-period risk premium is positively related to the short rate. The reverse is not necessarily true, however. We note that incorrect duration, whether too short or too long, could never explain the observation that the slope of the term structure gives wrong signals about the future path of interest rates.

Reuben Kessel (1965) ran regressions of forward-spot rate differences at the short end of the term structure on the short interest rate, and concluded that the forward rate premium was positively related to the short yield. Such a correlation could be taken to mean that long interest rates overreact to short rates. However, our work with more recent data shows that the effect of the short rate is, if anything, negative. Using monthly data from 1955:1 to 1979:8, and regressing the excess one-month return on a twenty-year bond over a one-month bill on the one-month bill rate, we find a coefficient of -0.479 with standard error 0.766. But the short rate has very little explanatory power (R^2 is only 0.001); it is rather the spread between long and short rates which explains excess holding returns, with an R^2 of 0.014 and a significant coefficient of 3.095, larger than unity. This is a reflection of the perverse behavior of the slope of the term structure in predicting future interest rates.[2]

There has been an uptrend in interest rates since Kessel's sample. This suggests an alternative overreaction or underreaction hypothesis that risk premia may be explained in terms of the difference between the short rate and a moving average or distributed lag of short rates. In fact, our results so far would seem to suggest just this, since as we noted in the introduction the long-short spread which explains excess returns is itself well described as a distributed lag on short rates. For our data, the estimated distributed lag places a weight of -0.805 on the current short rate and $+0.878$ on a five-year Almon cubic polynomial lag of short rates.[3] These coefficients lead us to expect that the risk premium is high when the short rate is low relative to recent experience. Nevertheless, when the excess return is regressed directly on current and lagged short rates, the point estimates are statistically insignificant with t-statistics of only about 0.1. This evidence is not inconsistent with rational forecasting in the 1955–79 period. We note however that when the sample is extended to the end of 1982, the coefficient on the current short rate becomes negative and significant at the 9 percent level, while the sum of the lag coefficients is positive and significant at the 7 percent level. This could be taken to imply that long rates have *underreacted* to short rates.

Another way to examine this issue is to conduct an ARIMA analysis of the behavior of short rates. Shiller's volatility analysis suggested that nonstationarity of interest rates might be necessary to justify the behavior of long rates; we assumed this conclusion and used monthly data over the period 1955–79 to estimate an ARIMA (1,1,1) process for the one-month bill rate. This specification has the important advantage of being time consistent, that is independent of the measurement interval. It implies that the long-short spread under the rational expectations theory of the term structure should be a function of current and lagged short rates, with the influence of lagged short rates declining geometrically at a rate equal to the *MA* parameter, and with the sum of the coefficients on lagged short rates equal to the negative of the coefficient on the current short rate. We found that the likelihood function was very flat, but was maximized by the model $(1-0.950L)\Delta R_t=(1-0.975L)u_t$. With these parameter values the rational expectations model implies that in the distributed lag equation for the spread the coefficient of the current short rate should be -0.47 and the sum of the lagged coefficients should be $+0.47$, with a very slow decay within the distributed lag. The Modigliani-Sutch distributed lag is roughly consistent with this, but has a more highly negative coefficient on the current short rate. This suggests that the rolling risk premium tends to be high when the short rate is low relative to its recent history.

When the short rate and its distributed lag are included in a regression together with the long-short spread, we find that both become

[2] We note here the curious fact that excess returns of common stock over short debt also bear a significant positive relation to the long-short spread (Campbell, 1983). This observation suggests that risk premia on different assets move together.

[3] The R^2 in this regression is 0.809; however, as Llad Phillips and John Pippenger (1979) have pointed out, the residuals from this type of equation are highly serially correlated so that "spurious correlation" may exaggerate the explanatory power of the regression.

significant, and the coefficient on the spread triples. The fitted values in this regression look something like a multiple of the residuals from the distributed lag equation for the spread, suggesting that the significance of the current and lagged short rates is due to the regression's trying to purge the long-short spread of the component which is explained by current and lagged short rates. When the fitted value and residual from the spread equation are included separately, only the residual is significant. It has a coefficient of 10.336 with a standard error of 3.440, while the fitted value has coefficient 1.388 and standard error 1.676. Splitting the spread into fit and residual more than doubles the R^2 to 0.032. It is also the residual which in the 1955–79 sample accounts for the violation, noted by Shiller, of variance restrictions on holding period yields. When the sample is extended to 1982, however, both the fitted value and the residual explain excess holding returns and violate the variance restrictions.

We see then that holding-period and rolling risk premia have if anything been negatively related to short rates, suggesting that long rates if anything have underreacted to short rates. If long rates had been a distributed lag on short rates, with a somewhat larger coefficient on the current short rate and smaller coefficients on lagged short rates, then excess holding returns on long bonds would have been less predictable than they in fact were. But this sort of underreaction was not primarily responsible for the failure of the expectations theory of the term structure. The independent movement of the long rate also violated the restrictions of the theory. In the 1955–79 period, it was that smaller part of the spread between long and short rates which was *not* explained by current and lagged short rates that caused excess volatility in holding period yields and destroyed the predictive power of the term structure.

REFERENCES

Ando, Albert and Kennickell, Arthur, "A Reappraisal of the Phillips Curve and the Term Structure of Interest Rates," unpublished paper, University of Pennsylvania, 1983.

Campbell, John Y., "Stock Returns, the Term Structure and Inflation," unpublished paper, Yale University, 1983.

Hansen, Lars Peter and Sargent, Thomas J., "Exact Linear Rational Expectations Models: Specification and Estimation," Staff Report, Federal Reserve Bank of Minneapolis, 1981.

Kessel, Reuben A., "The Cyclical Behavior of the Term Structure of Interest Rates," Occasional Paper No. 91, National Bureau of Economic Research, 1965.

Macaulay, Frederick, *Some Theoretical Problems Suggested by the Movements of Interest Rates, Stock Prices and Bond Yields in the United States Since 1856*, New York: National Bureau of Economic Research, 1938.

Mankiw, N. Gregory and Summers, Lawrence H., "Do Long-Term Interest Rates Overreact to Short-Term Interest Rates?," Council of Economic Advisers, 1983.

Modigliani, Franco and Shiller, Robert J., "Inflation, Rational Expectations and the Term Structure of Interest Rates," *Economica*, February 1973, *40*, 12–43.

______ **and Sutch, Richard C.,** "Debt Management and the Term Structure of Interest Rates: An Empirical Analysis of Recent Experience," *Journal of Political Economy*, August 1967, *75*, 569–89.

Phillips, Llad and Pippenger, John, "The Term Structure of Interest Rates in the MPS Model: Reality or Illusion?," *Journal of Money, Credit and Banking*, May 1979, *11*, 151–64.

Sargent, Thomas J., "A Note on Maximum Likelihood Estimation of the Rational Expectations Model of the Term Structure," *Journal of Monetary Economics*, January 1979, *5*, 133–43.

Shiller, Robert J., "The Volatility of Long-Term Interest Rates and Expectations Models of the Term Structure," *Journal of Political Economy*, December 1979, *87*, 1190–219.

______ **Campbell, John Y. and Schoenholtz, Kermit L.,** "Forward Rates and Future Policy: Interpreting the Term Structure of Interest Rates," *Brookings Papers on Economic Activity*, 1:1983, 173–217.

Tversky, Amos and Kahneman, Daniel, "Judgment Under Uncertainty: Heuristics and Biases," *Science*, September 1974, *185*, 1124–31.

Unanticipated Money and Interest Rates

By V. VANCE ROLEY AND CARL E. WALSH*

The relationship between money and interest rates is of fundamental importance to economic policymakers. In the absence of the liquidity trap, Keynesian theory emphasized the negative relationship between money and interest rates, thereby providing a role for the monetary authority to engage in countercyclical policy. Milton Friedman's (1968) accelerationist theory, however, questioned the efficacy of monetary policy activism. This theory predicted that increases in money growth would lead to identical increases in long-term interest rates. As a result, money growth and interest rates would be expected to exhibit positive correlation.

With the advent of rational expectations and efficient markets theories, the focus of this debate has shifted to the correlation of unanticipated money and interest rates. Despite the change in emphasis, the issues remain largely unchanged. In particular, do unanticipated increases in money lower real interest rates and hence stimulate economic activity? Or, alternatively, do unanticipated increases merely lead to similar rises in expected inflation, leaving real rates virtually constant?

Evidence on the relationship between unanticipated money and interest rates has been provided by two types of studies. First, several researchers have investigated the relationship using quarterly data. Among these, Frederic Mishkin (1981, 1982) found no evidence of negative correlation. Instead, his results indicated either positive or zero correlations for both short- and long-term interest rates.[1]

Second, a number of researchers have examined the effect of money announcement surprises on interest rates. Again, these studies have uniformly found positive correlations between surprises in announced money and both short- and long-term interest rates. In contrast to Mishkin, who very cautiously interprets his results, researchers examining the effects of weekly money announcements have typically preferred one of two common explanations for the positive correlation. One explanation advanced by a number of researchers is that the positive correlation resulted from short-run Federal Reserve policy (see, for example, Jacob Grossman, 1981; Thomas Urich and Paul Wachtel, 1981; Roley, 1983). Under this hypothesis, the Federal Reserve attempts to offset short-run deviations in money growth due to shifts in money demand. The other explanation, advanced by Bradford Cornell (1983), suggests that the positive correlation is due to associated changes in expected inflation.

The purpose of this paper is first to provide an interpretation of the positive correlation between unanticipated money and interest rates in terms of Federal Reserve policy objectives and operating procedures. Then, the correlation of unanticipated money and both short- and long-term interest rates is examined over weekly intervals. This empirical investigation combines some of the aspects of Mishkin's quarterly studies with those of the money announcement studies. In addition, the distinction between unpredicted and unperceived money, emphasized by Robert Barro and Zvi Hercowitz (1980), is also considered.

I. Federal Reserve Policy and Unanticipated Money

Unanticipated changes in money have been frequently interpreted as discretionary changes induced by the Federal Reserve. If institutional features of Federal Reserve pol-

*University of Washington, Seattle, WA 98195, and the National Bureau of Economic Research, and Princeton University and the National Bureau of Economic Research, respectively. We are grateful to Rick Troll for research assistance.

[1] In contrast to the positive correlation found in most studies, John Makin's (1983) results suggest negative correlation. His use of averaged interest rate data, however, may account for at least some of the difference.

icymaking are taken into account, however, this interpretation is only one of several possible alternatives. To consider the potential sources of unanticipated money, the role of Federal Reserve policy in general and monetary targets in particular should be examined.

As is well known, the Federal Reserve has targets for a set of monetary and credit aggregates for both annual and shorter periods. A set of annual targets is announced for each calendar year in conjunction with the Humphrey-Hawkins Act. While the Federal Reserve has the opportunity to change these long-run targets at a midyear review, this opportunity has seldom been used. Moreover, at the midyear review, the Federal Reserve is required to specify preliminary annual targets for the following calendar year. Thus, explicit annual targets, along with statements by Federal Reserve officials pertaining to trend monetary growth, enable the public to infer the long-run goals of current monetary policy.

Short-run monetary targets are set throughout the year at meetings of the Federal Open arket Committee (FOMC). These targets are usually set such that future money growth will eventually fall within the annual target ranges. In contrast to the annual targets, however, current short-run objectives are in principal unknown to the public until around the time of the next FOMC meeting. Thus, the public must assess the Federal Reserve's short-run objectives on the basis of observed policy actions.

A third relevant feature of monetary policy concerns the time at which monetary information becomes available to the Federal Reserve. Because of reporting lags, data on the narrowly defined money stock are available only shortly before the Federal Reserve's weekly money announcements. As a result, contemporaneous money is unknown to the Federal Reserve in any given statement week.

What do these institutional features imply about unanticipated money growth? One implication is that there are three potential sources of money surprises. First, unanticipated money may reflect changes in the Federal Reserve's long-run monetary targets. Such changes, however, would most likely reflect changes within the stated target ranges. Second, money surprises may either result from the public's misconception of short-run monetary policy objectives or unanticipated changes in these objectives. Finally, unanticipated money growth may reflect weekly fluctuations unknown to both the public and the Federal Reserve.

Depending on which of these sources is most prevalent, the positive correlation between money surprises and interest rates may be interpreted several different ways. If unanticipated money results from discretionary changes in the Federal Reserve's long-run targets, or trend money growth in general, such unanticipated changes would be expected to be correlated with changes in expected inflation. In turn, the response of long-term interest rates, and perhaps to some extent short-term interest rates as well, would be due to changes in expected inflation.

If, however, unanticipated money does not reflect changes in long-run policy objectives, it is difficult to ascribe the response of interest rates to changes in expected inflation. Instead, the positive correlation between unanticipated money and interest rates may be due to the Federal Reserve's desire to offset short-run deviations in money growth. Again, such deviations only become apparent to the Federal Reserve after the statement week in which they occur. In this case, we have demonstrated in an earlier paper (1983) that the observed positive correlation between unanticipated announced changes in money and both short- and long-term interest rates may be explained in a model incorporating such policy responses. Our evidence further suggested that the Federal Reserve offsets short-run deviations in money within one year, implying that unanticipated money does not relate to changes in trend money growth.

In this same study, we also considered the possible effects of different Federal Reserve operating procedures on the correlation between unanticipated announced changes in money and interest rates. The larger positive correlation found after October 1979—when the Federal Reserve shifted from a federal funds rate to a nonborrowed reserve operating procedure—could be explained by two factors. In particular, following Walsh (forthcoming), greater volatility in short-term interest rates may have led to a reduction in the interest rate elasticity of the demand for money. Thus, if shifts in money demand are persistent, larger movements in interest rates would be required to return money to its long-run target. Moreover, money announcements also provide information about the aggregate demand for required reserves because of lagged reserve accounting. With the adoption of the reserves aggregate operating procedure in October 1979, this factor also helps to explain the increased positive response of short-term interest rates to positive surprises in announced money.

II. Empirical Evidence

As mentioned, evidence suggesting positive, or at least nonnegative, correlation between unanticipated money and interest rates has been provided by both quarterly studies and investigations of the response to weekly money announcements. The empirical investigation reported here attempts to combine some of the aspects of these different approaches. In particular, movements in interest rates are measured over an entire statement week. This enables the role of the money announcement occurring in a statement week to be determined. The hypothesis underlying this approach is that the money announcement is used to revise the estimate of the current week's money stock. The correlation of this revision, as well as the expectational error that remains, with both short- and long-term interest rates is then empirically examined. The amount of unperceived money in the current statement week is further decomposed into the forecast error associated with the money stock as first announced and as eventually revised.

The basic specification used to examine the response of interest rates to unanticipated money may be represented as

$$(1)\quad \Delta R_t = b_0 + b_1 \cdot UM_t + b_4 \cdot M_t^e + u_t,$$

where ΔR_t is the change in either the three-month Treasury bill yield ($R3M$) or the ten-year constant-maturity Treasury security yield ($R10Y$) from 3:30 P.M. on Wednesday of the previous statement week $(t-1)$ to 3:30 P.M. on Wednesday of the current statement week (t); UM_t is the difference between the log of the actual level of the narrowly defined money stock in week t and its expected level as of the end of week $t-1$; M_t^e is the log of the expected level of the money stock in week t as of week $t-1$; u_t is a random error term uncorrelated with all publicly available information in week $t-1$; and the b_i are coefficients to be estimated.[2] Under the hypothesis of rational expectations, the effect of anticipated money equals zero ($b_4 = 0$).

Unanticipated money (UM_t), defined to equal the log of week t's actual money stock minus the log of the market's expectation prior to the Friday announcement, can be decomposed into three separate factors:

$$(2)\quad UM_t = M_t - M_t^b = (M_t - M_t^p) + (M_t^p - M_t^a) + (M_t^a - M_t^b),$$

where M_t is the log of the actual narrowly defined money stock (as of October 1983), M_t^b is the expectation of M_t before the Friday announcement, M_t^a is the expectation after the announcement, and M_t^p is the initially announced value of M_t (M_t^p is the figure released on Friday of week $t+2$).

To form the expectation of the current week's narrowly defined money stock, a simple autoregressive process is used. However, to take advantage of available survey data on the level of the money stock to be announced in the current statement week, the current week's expected money stock before the announcement is taken as the fitted value of

$$(3)\quad M_t^p = a_0 + a_1 \cdot M_{t-2}^s + \sum_{i=2}^{m} a_i \cdot M_{t-i-1}^p + v_t,$$

where M_{t-2}^s is a survey for the money announcement in week t; v_t is a random error

[2] Following James Pesando (1979), the random walk model is used to represent weekly movements in interest rates. This approximation is likely to be good for the ten-year yield, but it may be somewhat inadequate for the three-month yield.

term; and the a_i are coefficients to be estimated.[3] The expectation of the current week's money stock after the money announcement is then taken as the fitted value of

$$(4) \quad M_t^p = a_0' + a_1' \cdot M_{t-2}^p + \sum_{i=2}^{m} a_i' \cdot M_{t-i-1}^p + e_t,$$

where e_t is a random error term and the other variables are defined as before. Thus, v_t is used to represent unanticipated money in week t as initially announced in week $t+2$, and e_t is used as the measure of unanticipated money after M_{t-2}^p becomes known.

Equation (2) can now be rewritten as

$$(2') \quad UM_t = (M_t - M_t^p) + e_t + (v_t - e_t).$$

This formulation highlights data revisions, following Barro and Hercowitz, and expectations revisions in response to the new information about M_{t-2}^p available during week t. Note that $v_t - e_t$ simply represents the revision in the estimate of the current week's money stock due to the weekly money announcement. If $a_i = a_i'$, this measure is proportional to the money announcement surprise.[4] In contrast, both e_t and $(M_t - M_t^p)$ are unperceived in the aggregate throughout the current statement week. To allow different responses to the different components of unanticipated money, the specification examined empirically is

$$(1') \quad \Delta R_t = b_0 + b_1 \cdot (v_t - e_t) + b_2 \cdot e_t + b_3 \cdot (M_t - M_t^p) + b_4 \cdot M_t^e + u_t.$$

To further allow for different interest rate response due to different Federal Reserve operating procedures, equation (1') is estimated separately for the pre- and post-October 1979 periods. For the pre-October 1979 period—beginning with the statement week of September 29, 1977 and ending with the statement week of October 5, 1979—the estimation results are

$$(5) \quad \Delta R3M_t = \underset{(5.745)}{-9.850} - \underset{(14.813)}{17.567}(v_t - e_t) + \underset{(4.764)}{4.313}e_t - \underset{(3.096)}{2.403}(M_t - M_t^p) + \underset{(0.981)}{1.692}M_t^e + u_t$$

$$\bar{R}^2 = .019, \; S.E. = .300, \; D\text{-}W = 2.056;$$

$$(6) \quad \Delta R10Y_t = \underset{(2.023)}{-1.855} - \underset{(5.229)}{4.182}(v_t - e_t) + \underset{(1.682)}{1.857}e_t - \underset{(1.093)}{0.063}(M_t - M_t^p) + \underset{(0.346)}{0.320}M_t^e + u_t$$

$$\bar{R}^2 = -.009, \; S.E. = .106, \; D\text{-}W = 1.728;$$

where standard errors of estimated coefficients are in parentheses, $S.E.$ is the standard error, $\bar{R}^2$ is the multiple correlation coefficient corrected for degrees of freedom, and $D\text{-}W$ is the Durbin-Watson statistic. The estimation results for both the three-month and ten-year yields fail to indicate a significant response to any of the categories of unanticipated or anticipated money. In contrast to weekly money announcement studies, the impact of this new information $(v_t - e_t)$ is insignificant. However, this result is not totally unexpected in light of the small intervals used previously to obtain estimated responses to money announcement surprises. Similarly, measures of unperceived money are not statistically significant in either regression.

[3] The survey data are collected by Money Market Services, Inc. As discussed below, separate autoregressions were estimated for the pre- and post-October 1979 periods. For all autoregressions, however, the last value of the money stock included is M_{t-5}. Also, despite the notation, lagged values of the money stock include revisions known as of the beginning of week t.

[4] The hypothesis that $a_i = a_i'$ in both the pre- and post-October 1979 periods could not be rejected at the 5 percent significance level.

For the post-October 1979 period—beginning with the statement week of October 8, 1979 and ending with the statement week of October 13, 1982—the estimation results are[5]

$$(5') \quad \Delta R3M_t = \underset{(7.228)}{2.802} + \underset{(16.951)}{38.960^*}(v_t - e_t) + \underset{(8.132)}{47.062^*}e_t + \underset{(10.366)}{10.198}(M_t - M_t^p) - \underset{(1.200)}{0.479}M_t^e + u_t$$

$$\bar{R}^2 = .227,\ S.E. = .714,\ D\text{-}W = 2.027;$$

$$(6) \quad \Delta R10Y_t = \underset{(3.783)}{5.410} - \underset{(8.799)}{6.607}(v_t - e_t) + \underset{(4.222)}{16.649^*}e_t + \underset{(5.381)}{0.896}(M_t - M_t^p) - \underset{(0.623)}{0.897}M_t^e + u_t$$

$$\bar{R}^2 = .128,\ S.E. = .370,\ D\text{-}W = 1.788;$$

where asterisks indicate statistical significance at the 5 percent level. These estimation results differ sharply from those obtained in the pre-October 1979 period. First, the money announcement surprise significantly affects the three-month yield. The estimated coefficient implies that a 1 percent money surprise causes the three-month yield to increase by almost 39 basis points. Second, changes in both the three-month and ten-year yields are significantly correlated with the expectational error remaining after the current week's money announcement, but not with the error associated with subsequent data revisions. In the case of the three-month yield, for example, a 1 percent positive surprise results in over a 47 basis points increase.

Because e_t is unperceived during week t, it is not likely that its effect on interest rates can be attributed to any induced revision of expected inflation. This positive response can, however, be interpreted in terms of short-run reserve adjustment by banks to weekly fluctuations in private sector money demand. An upward shift in money demand can exert a contemporaneous upward effect on interest rates, even under lagged reserve accounting, as individual banks begin to adjust their reserve position. Prior to October 1979, the Federal Reserve would have prevented rates from moving.

To summarize, no significant correlation between interest rates and unanticipated money was found in the pre-October 1979 period. Not even announced money surprises were found to significantly affect interest rates over weekly periods. This result is, however, probably due to the substantially larger variance in the change in interest rates when moving from daily to weekly intervals. In the post-October 1979 period, announced money surprises nevertheless had a significant positive correlation with the three-month yield. Moreover, another component of unanticipated money—measuring the expectational error remaining after the current week's money announcement—was statistically significant and positively correlated with changes in both short- and long-term interest rates. The most plausible explanation of this response is again based on bank reserve adjustment and the anticipated reaction of the Federal Reserve to short-run deviations in money growth.

The nature of money market shocks which characterize the post-October 1979 sample period has led to a positive correlation, as Friedman predicted, between money shocks and interest rates. This correlation, however, has little, if anything, to do with expected inflation.

REFERENCES

Barro, Robert J. and Hercowitz, Zvi, "Money Stock Revisions and Unanticipated Mon-

[5] The estimation period ends in October 1982 due to the Federal Reserve's apparent deemphasis of the reserves-aggregate operating procedure around that time.

ey Growth," *Journal of Monetary Economics*, April 1980, *6*, 257–67.

Cornell, Bradford, "Money Supply Announcements and Interest Rates: Another View," *Journal of Business*, January 1983, *56*, 1–24.

Friedman, Milton, "The Role of Monetary Policy," *American Economic Review Proceedings*, May 1968, *58*, 1–17.

Grossman, Jacob, "The "Rationality" of Money Supply Expectations and the Short-Run Response of Interest Rates to Monetary Surprises," *Journal of Money, Credit and Banking*, November 1981, *13*, 409–24.

Makin, John H., "Real Interest, Money Surprises, Anticipated Inflation and Fiscal Deficits," *Review of Economics and Statistics*, August 1983, *65*, 374–84.

Mishkin, Frederic S., "Monetary Policy and Long-Term Interest Rates: An Efficient Markets Approach," *Journal of Monetary Economics*, January 1981, *7*, 29–55.

______, "Monetary Policy and Short-Term Interest Rates: An Efficient Markets-Rational Expectations Approach," *Journal of Finance*, March 1982, *37*, 63–72.

Pesando, James E., "On the Random Walk Characteristics of Short- and Long-Term Interest Rates in an Efficient Market," *Journal of Money, Credit and Banking*, November 1979, *11*, 455–66.

Roley, V. Vance, "The Response of Short-Term Interest Rates to Weekly Money Announcements," *Journal of Money, Credit and Banking*, August 1983, *15*, 344–54.

______ **and Walsh, Carl E.,** "Monetary Policy Regimes, Expected Inflation, and the Response of Interest Rates to Money Announcements," Working Paper No. 1181, National Bureau of Economic Research, 1983.

Urich, Thomas J. and Wachtel, Paul, "Market Responses to the Weekly Money Supply Announcements in the 1970's," *Journal of Finance*, December 1981, *36*, 1063–72.

Walsh, Carl E., "Interest Rate Volatility and Monetary Policy," *Journal of Money, Credit and Banking*, forthcoming.

Interaction Between Fiscal and Monetary Policy and the Real Rate of Interest

By ROBERT ANDERSON, ALBERT ANDO, AND JARED ENZLER*

One of the consequences of the economic policies pursued since 1981 is the prospect of a continued large federal deficit combined with a high level of the market rate of interest for at least several years to come in the United States. This is a radical departure from the pattern that prevailed in the United States during the period after World War II. Except for cyclical and temporary variations, federal debt (net of holdings by government agencies) as a proportion of net national product declined steadily between 1945 and 1980, while the (*ex post*) real rate of interest remained quite low.[1] This fundamental shift in the economic policy since 1981 has important consequences for both the U.S. economy and the world economy as a whole.

This paper provides a quantitative analysis of these consequences. In order to examine such a complex problem, a specific model of the economy with numerical estimates of its parameters is needed. We have chosen to work with the one most familiar to us, namely, the MPS econometric model of the United States. The nature of this model requires us to conduct the short-run dynamic and medium- to long-run analysis together, but, given the space available, we only report on the latter.

In its description of aggregate demand, the MPS model is essentially Keynesian, and thus it is possible with many qualifications, to summarize the features of its aggregate demand functions in the form of an *I-S* curve.

On the other hand, the model also describes the process of capital accumulation and its effects on the productivity of labor in some detail. The capacity of the economy to produce output, although it is defined and measured in a limited sense, does play a critical role in determining price dynamics and the productivity per unit of labor input.

The rate of change in the wage rate is determined by an expectations augmented Phillips curve. The value-added prices for output are set in accordance with a markup behavior resulting from the cost minimization by firms faced with oligopoly markets for output. Final goods prices are the appropriately weighted averages of value-added prices and raw material prices. The interaction among these three components determines the dynamic behavior of prices and wages, given those for raw materials. Our estimates of parameters do imply the existence of a "natural rate of unemployment," which we prefer to call by the more neutral name "nonaccelerating rate" of unemployment (*NARU*). The *NARU* is consistent with the maintainance of any constant rate of inflation; when the rate of unemployment is below *NARU*, the rate of inflation tends to accelerate while when it is above *NARU*, the inflation rate tends to deccelerate. Exogenous movements in raw materials prices can also affect the inflation rate, as in the case of oil shocks in the 1970's (supply shocks), but the resulting acceleration of the inflation rate cannot be sustained if the actual unemployment rate is in the appropriate relation to *NARU*. It should be noted that we attach no normative significance to *NARU*. Indeed, we believe one of the most important public policy questions facing the OECD countries is how to design policies to lower *NARU*, although the subject is outside the scope of this paper. In our model, *NARU* responds over the long run to the rate of growth of

*Anderson and Enzler: Board of Governors of the Federal Reserve System; Ando: Department of Economics, University of Pennsylvania, Philadelphia, PA 19104. Views expressed in this paper are our own and do not necessarily reflect those of the Board of Governors, its staff, or any other organization.

[1] The exception is the rate of return on equities for several years after the war, and again after 1973.

productivity per man-hour, an empirical result with which we are not completely comfortable.

Monetary policy actions affect prices by first affecting the real rate of interest and hence the aggregate demand, and eventually the rate of unemployment. The money supply has no direct effect on price movements outside of this basic channel, a point which may be subject to dispute among researchers. Ando and Arthur Kennickell (1983) addressed this issue in detail in an earlier paper.

In this model, expectations by economic agents of the future course of endogenous variables do play an essential role. The formation of expectations is largely modelled as adaptive, and thus the model is subject to what has become known as the "Lucas critique" of econometric policy evaluation. For further discussion of this issue, see Anderson and Enzler (1983). For the purposes of the present paper, however, this question is not critical because here we are interested in the long-run effects of various fiscal policy options on steady-state growth paths, along which expectations become self-fulfilling in this model.

While the model is Keynesian in its description of the short-run aggregate demand sectors, it could follow a steady-state growth path in which the rate of unemployment is at *NARU* and whose characteristics are largely defined by supply considerations: labor force growth, technological progress, capital formation, and so on. Such a path is not unique for the model, nor any of them necessarily stable dynamically. Both the position of the path and its dynamic stability depend critically on the nature of fiscal and monetary policies followed by the government. In the remainder of this paper, by "long-run" impact of a policy change, we mean the effects of the policy change on the pattern and characteristics of the steady-state growth path of the model.

In this model, government debts are treated as assets owned by the private sector, which capitalizes future tax payments only imperfectly. Indeed, the relationship between current deficits and future taxes is not at all simple, though it does exist and plays an important role as shown later in our analysis. Fiscal policy, therefore, does have long-run effects on the economy, both because the size of the government debt has effects on the capital formation given the savings behavior, and also because a number of tax and subsidy provisions create wedges between prices received by suppliers and those paid by purchasers, especially between the net user cost of capital and the market rate of interest.

Monetary policy, on the other hand, has only a very limited effect on the long-run steady-state path of the economy. It, of course, determines the level and the rate of change of prices, and the model is not superneutral reflecting, among other things, the fact that most tax provisions are defined in nominal terms. Except for these tax provisions, and for external financial relations, monetary policy cannot affect real variables on the steady state growth path. It is, however, capable of exerting powerful impacts on the real economy in the short run by first affecting the real rate of interest and through it all other real variables in the economy.

Taking advantage of these features of monetary and fiscal policies, we propose to analyze the long-run consequences of alternative fiscal policy options by first defining alternative fixed fiscal policies, and then running dynamic simulations of our model for each of these fiscal policy options with monetary policy rules designed to stabilize the economy represented by the model as long as necessary to observe steady-state growth paths corresponding to each of these fiscal policy options.

For the MPS model, we have been able to construct a class of monetary policy feedback rules which set the federal funds rate in response to the rate of inflation, to deviations of the actual unemployment rate from *NARU*, to deviations of output from the capacity output, and deviations of the actual rate of growth of output from the rate of growth implied by rates of growth of labor force and of productivity per man-hour. These feedback rules are capable of pushing the economy within a small distance of *NARU* while also achieving the target rate of inflation within a few years starting from any initial conditions that are not too outrageous.

The nature and the functioning of such a rule is quite complex, and we shall defer its discussion to another, longer paper. We merely note here that, given the nature of the dynamic structure of our model, rules calling for a strictly fixed rate of growth of the money supply, or a rigidly fixed federal fund rate, are destabilizing rather than stabilizing. When the economy converges to a steady-state growth path with *NARU* under the guidance of such a monetary policy feedback rule, the rule will have found the particular real rate of interest consistent with this growth path.

In order to carry out our program of generating steady-state growth paths associated with several alternative fiscal policies with the aid of a stabilizing monetary policy rule, we must set fiscal policy parameters that are both consistent and feasible. An important condition that must be observed is the stability of the government's budget position. In order to clarify this point, it is helpful to express the government budget constraint in the form given by

$$\dot{d} = d[r(1-\tau') - (\dot{P}/P + \dot{X}/X)] + (g-\tau), \tag{1}$$

where d, g, and τ are, respectively, ratios of total net government debt (including Federal Reserve holdings but excluding holdings of other federal government agencies), total federal government expenditures excluding interest payments on government debt, and federal tax revenues from all sources, to nominal *GNP*; τ' is the tax recovery rate on the federal interest payments, r is the nominal rate of interest on federal debt, and X and P are real *GNP* and its deflator. Dots above symbols denote time derivatives.

A glance at this expression makes clear that government finance would face a serious instability problem when the quantity in brackets is positive, that is, when the nominal rate of interest on government debt is greater than the rate of growth of nominal *GNP*. If this is the case, then government debt as a proportion of nominal *GNP* will grow without bound unless there is a sufficiently large surplus in the current account, that is, $g-\tau$ is sufficiently negative, to offset it.

To avoid this instability problem, in specifying fiscal policy alternatives we proceed as follows. We first set parameter values for all government revenues and expenditures other than personal income taxes, as called for by the specification of a particular fiscal alternative. Next, we set the target ratio for the value of total net national debt to nominal *GNP*. We then let the average effective rate of federal personal income tax be endogenously set by a rule, so that $\dot{d}$ in equation (1) always takes on the value which reduces the gap between the actual value of d and its target. When d reaches its target, the effective rate is set to insure that $\dot{d} = 0$. The rule determining the effective rate is designed so that the relative rate structure of the personal income tax over income distribution remains unchanged.

When the domestic rate of interest varies for domestic reasons, the exchange rate, capital account balances, and the condition of the current account all must be affected. In the extreme case, it is possible that additional borrowing by government is financed almost entirely by a large net inflow of foreign capital, leaving domestic conditions relatively unaffected.

There are two reasons why the external sector must be monitored closely. The relationship of the stock of net foreign capital holdings to some scale variable, say net domestic product, can become unstable precisely for the same reason as the relationship of national debt to net national product may become unstable, as described above. Since we shall consider cases in which the real rate of interest is substantially larger than the maintainable rate of growth, we must have a policy mechanism in place to keep foreign debts within reasonable limits in such situations.[2]

The second reason is related to the complex dynamic adjustment of the external balances and its interaction with domestic stabilization processes. An increase in the U.S. rate of interest creates an imbalance in the

[2] The difficulties faced by some *LDC*s in managing their international debt since 1980 can be more easily understood in this framework.

capital account initially, but, except for capital gains and losses associated with changes in the exchange rate, the net stock of capital accounts can change only through movements in the current account. Since the current account, except for interest payments, does not respond to movements in the rate of interest directly, it is the exchange rate that must move to offset the initial change in the rate of interest so that the capital accounts will remain in balance. The capital accounts, however, depend on the expected *rate of change* in the exchange rate, not its level. On the other hand, imports and exports respond to the level of the exchange rate. The exchange rate will eventually settle with a higher rate for the dollar than its initial rate, so that U.S. net foreign investment will be sufficiently lower than before to accommodate the needed change in capital accounts. This means that the exchange rate must initially overreact sufficiently so that its eventual higher value is also consistent with the expectation of a fall in its value. These movements of the exchange rate can be quite large, and they can in turn cause substantial oscillatory movements in domestic prices, and hence, complications for domestic stabilization policies. To control these interactions, we have in effect adapted a "dirty float" regime.

The size and speed of all these reactions, of course, depend not only on the relevant elasticities involved, but also on the reactions of foreign economic magnitudes to movements of domestic variables, especially the response of foreign interest rates to variations in the domestic rate of interest. The MPS model contains very rudimentary functions describing the responses of these foreign economic magnitudes to domestic variables. In the case of foreign interest rate reactions to the domestic rate, recognizing that there are political elements in such reactions, we have left the flexibility in the model for the user to specify the strength of these reactions. For the analysis in this paper, we have considered two cases: one in which foreign rates of interest move completely parallel to the domestic rate, so that net capital accounts remain unaffected by change in the domestic rate of interest; and a second case in which foreign rates respond strongly but not completely to the domestic rate, so that the capital accounts are moderately affected by changes in the domestic rate.

As stated before, the vehicle of our analysis is to run the MPS model for each of fiscal policy choices under a monetary stabilization rule as long as necessary to arrive at a steady-state growth path. The stabilization rule is set so that the steady-state inflation rate is 5 percent per year in all of our simulations. The *NARU* is approximately 6.7 percent. Cyclical fluctuations die down reasonably soon, usually within ten years, but it takes much longer for various stocks to approach equilibrium relations to each other. To be on the safe side, we have chosen fifty years (200 quarters) after the start of the simulation as the point where the system has, for all practical purposes, attained its stationary state.

Needless to say, even though we figuratively start our simulations from the present quarter, the steady-state growth paths described below have nothing to do with forecasts of the actual economy for the year 2033. These steady-state paths are analytical concepts computed in order first to see what the model says about long-run consequences of alternative fiscal policies, and second to help us understand the model's deficiencies for this type of analysis.

Let us consider first the case in which the foreign response is "mild" in the sense that, when the domestic rate of interest rises, the net ownership of foreign assets by U.S. residents declines moderately, and the structure of government expenditures and taxes remain as they are now except for the effective rate of the personal income tax, which is used as the instrument to achieve the debt target. In simulations I, II, and III, the debt target is set at zero, 30, and 100 percent of nominal *GNP*. As the debt target moves, the equilibrium real rate of interest (as measured by the federal funds rate less the rate of inflation) moves 1.6, 2.7, and 6.1 percent, respectively. In simulation I, the rate of growth of nominal *GNP* is clearly greater than the nominal rate of interest, and the government budget is in balance; in simulation II, the rate of growth is marginally larger than the rate of interest, and the government runs a small deficit as defined in equation (1) above (deficits in the standard national income account definition less government interest payments plus tax recovery on government interest payments); and in simulation III, the rate of growth is clearly smaller than the rate of interest, and the government must run a substantial surplus by the same definition. The corresponding effective rate of income tax is highest for simulation I, lowest for simulation II, and it is between these two values for simulation III. This is an illustration of the points that a larger national debt does not necessarily imply higher taxes in a growing economy with an infinite horizon, and that individuals' perception that government bonds are as much asset items as other private obligations may be perfectly consistent with the outcome generated by the economic system.

There are, in this economy, four ultimate holders of assets, namely, households, foreigners, state and local governments, and federal government. From simulation I to simulation II to simulation III, the federal government is increasing its indebtedness, and other sectors are correspondingly increasing their net ownership of federal obligations. State and local governments absorb a small quantity, foreigners are increasing their net ownership of the U.S. assets from negative 9.5 percent to negative 3.5 percent, and then to positive 13.4 percent of the U.S. Gross National Product. Since the federal government increases its indebtedness by 100 percent of *GNP* from simulation I to simulation III, the household sector must be induced to hold most of it, approximately 75 percent of *GNP* (100 less 22.9 percent absorbed by foreigners and 2 or 3 percent absorbed by state and local governments).

Total net worth of households, as a percentage of *GNP*, remains roughly constant from simulation I to stimulation III, at roughly 300 percent. Households must therefore reduce their holdings of other assets in order to acquire government debts. A part of the reduction comes from residential housing and consumer durables, and another part is a reduction in the relative value of land. A large part, however, is a reduction in consumers' indirect holdings of productive capital, in the form of equities and debts of corporations.

Our simulation results indicate that the market value of these holdings is indeed reduced by a substantial proportion, by a little less than a third, from simulation I to simulation III. By constrast, the physical volume of productive capital held by corporations (producers' structure, durable equipment and inventories) declines by some 15 percent. This discrepancy is partly due to the fact that the complex provisions of the corporate profit tax laws create significant disparity between the market value of investment goods and their cost to corporations, a disparity that is itself a function, among other things, of the rate of inflation and the level of the real rate of interest.

As the real rate of interest rises from 1.6 to 6.1 percent, and the level of capital stock per worker is lowered from simulation I to simulation III, productivity per man-hour and total output in the economy are reduced. The *GNP* in constant dollars is indeed some 5 percent smaller in simulation III relative to that in simulation I. The level of consumption, however, does not go down much from simulation I to simulation III. This illustrates the point that, when the underlying production function is well behaved, as it is in this model, it is possible that within limits, variations in the capital output ratio have relatively small effects on consumption. This can happen when the system is, even though by an accident, near the Golden Rule conditions.

We may note at this point that, if a higher real rate of interest in the United States did not attract additional foreign investment in the United States, the effects of a higher government debt-*GNP* ratio is qualitatively the same but quantitatively considerably more severe.

The last question to which we wish to address ourselves is to what extent various special provisions in the tax law for encouraging investment and capital formation, such as investment tax credit, accelerated depreciation allowances, and the deductibility of interest payments on business loans and on mortgages are serving to raise the market rate of interest in the U.S. economy, and having other effects that are often overlooked. In order to deal with this question, we modified the tax provisions in our model: specifically we (*i*) eliminated the investment credit from the tax laws; (*ii*) replaced current depreciation allowance schedules for investment goods for businesses by the system in which they can expense capital goods purchases, with carryover provisions at the market rate of interest; (*iii*) eliminated the deductibility of interest payments for business taxes; (*iv*) eliminated the deductibility of interest payments on mortgates for residential houses.

With these modifications implemented in our model, we then ran the identical simulations II and III discussed earlier, that is, with target national debt-*GNP* ratio of 30 and 100 percent. We shall refer to these simulation results with the modified tax structure as simulations II′ and III′, respectively.

The equilibrium real rate of interest falls from 2.7 percent in simulation II to 1.4 percent in simulation II′, and from 6.1 percent in simulation III to 4.6 percent in simulation III′. These are very significant and large differences.

Effects of these provisions on capital accumulation are, according to our model, quite sizeable. Physical stocks of capital are all smaller in simulation II′ relative to those in simulation II and also in simulation III′ relative to those in simulation III, although the size of responses varies from one particular type of stock to another. On average, the decline is some 15 percent Coincidentally, the stocks of productive capital in II′ are very similar to those in III. Simulation II′, however, represents a much healthier economy than III because there are far fewer allocational distortions both within the economy and in the relation of the U.S. economy with those of other countries.

We wish to point out again that the policies resulting in simulation III, the ones which most closely resemble the ones currently in place, generate a rate of interest clearly above the rate of growth, leading to serious instabilities for government finance and for international debt balances. Unlike the simulation experiments, the actual political context might not permit controls to contain these instabilities, which would then become quite serious.

Results such as these are essential for evaluating consequences of major shifts in fiscal policies. They cannot be generated without a detailed model, and therefore, they are model specific and subject to the defects of the model used. Although we do not believe that this limitation is so serious as to make these results meaningless, we would welcome detailed discussions of the model features and suggestions for improvements in order to make our results more representative of the real economy.

REFERENCES

Anderson, Robert and Enzler, Jared, "Designing Stabilization Policy Reaction Functions," unpublished manuscript, 1983.

Ando, Albert and Kennickell, Arthur, "'Failure' of Keynesian Economics and 'Direct' Effects of the Money Supply: A Fact or a Fiction?," unpublished manuscript, March 1983.

Blinder, A. S., "Reaganomics and Growth: The Message in the Models," paper presented at the Conference on Reagan Economic Programs and Long-Term Growth, September 22–23, 1983.

Practical Implications of Game Theoretic Models of *R&D*

By JENNIFER F. REINGANUM*

The purpose of this paper is to survey recent game theoretic models of research and development, and to ask whether they yield practical implications or testable hypotheses. The papers I will be discussing have been written or published within the past five years; they have a good deal in common, relying upon similar assumptions and building upon each other. Because of this, the literature surveyed here may seem narrow, and some relevant work has probably been inadvertently omitted. I apologize for these omissions. For want of space, I will not discuss the normative conclusions of this literature.[1]

I. Why Game Theory?

As students of industrial organization, we cannot ignore interactions among the agents we study. Positive industrial organization is the study of business policy and strategy. Modern noncooperative game theory is a language of strategy and equilibrium; that is, it provides an equilibrium framework in which to examine individuals' strategic behavior. Recent advances, for instance the theory of supergames (James Friedman, 1971) and the notions of perfect equilibrium (Reinhard Selten, 1975) and sequential rationality (David Kreps and Robert Wilson, 1982), have made game theory an even more powerful tool for examining controversial issues in industrial organization. All models must postulate the behavior of some agents in the model; a game theoretic model must, in addition, impose certain consistency checks, or equilibrium conditions, upon this postulated behavior. Within the confines of the game theoretic paradigm, there are still many alternative modeling choices regarding, for example, informational assumptions and timing conventions. Thus the paradigm is capable of generating a wide range of equilibrium behavior. As with any theory, the ultimate appeal for validation or vitiation is to empirical testing. I will isolate and discuss several implications, most of them controversial, from the recent literature. This seems to be a useful first step toward the goal of empirical testing.[2]

II. The Implications of Recent Models

Because most of the papers discussed below analyze and extend a single basic model, I provide a brief description of this model. Partha Dasgupta and Joseph Stiglitz (1980) and Glenn Loury (1979) employ a model of stochastic invention in which the probability of success by firm i by a given time, t, is an exponential function. That is, if t_i represents firm i's (random) success date, then $Pr(t_i \leq t) = 1 - e^{-h_i t}$, where h_i is the "hazard rate," or conditional probability density of success, given no success to date. The choice variable

*Division of Humanities and Social Science, California Institute of Technology, Pasadena, CA 91125. I thank Therese Flaherty, Nancy Gallini, Avinash Dixit, and Alan Schwartz for helpful comments and suggestions. The financial support of the National Science Foundation is gratefully acknowledged.

[1] For a more complete survey of the previous empirical and theoretical literature on this subject, see Morton Kamien and Nancy Schwartz (1982). For a detailed account of previous empirical work and a discussion of appropriate measurement and methodology, see David Grether (1974).

[2] Although previous empirical work is suggestive, most of it has not been carried out with a specific behavioral model of the firm in mind. Moreover, it has tested hypotheses which were couched in much more aggregated terms than those to be discussed below.

for each firm i is a lump sum expenditure x_i at time $t = 0$, which implies a hazard rate of $h_i = h(x_i)$. With this specification, the expected time till success for firm i is the reciprocal of the hazard rate; $E(t_i) = 1/h(x_i)$. The innovation "production function" $h(x)$ is allowed to have initial increasing returns to scale, but eventually decreasing returns set in. Patent protection is assumed perfect, firms are identical, and no further innovation is anticipated. This problem is modeled as a simultaneous-move game, and the equilibrium concept is Nash equilibrium in investment strategies. Tom Lee and Louis Wilde (1980) modified this formulation by assuming that the investment is a *flow* cost, rather than a lump sum payment at the initial date. That is, firm i pays at the rate x_i, but only until someone succeeds. They maintain all of the remaining assumptions of the model. The first three implications turn on this difference in the specification of costs.

1. The amount invested by an individual firm decreases with the number of firms engaging in *R&D*; however, aggregate industry investment increases with the number of firms.

1′. The *rate* of investment by an individual firm increases with the number of firms engaging in *R&D*; a fortiori, the aggregate industry investment rate increases with the number of firms.

Using the Dasgupta-Stiglitz and Loury fixed cost model, one can conclude that the equilibrium amount invested by any one firm decreases with the number of firms engaged in research and development. Despite this, an increase in the number of firms engaged in *R&D* results in an increase in aggregate investment. From the Lee-Wilde flow cost model, one can deduce that the equilibrium rate of expenditure per firm increases with an increase in the number of firms; a fortiori, aggregate investment increases with the number of firms. The intuition behind these conclusions is simple. In the Dasgupta-Stiglitz and Loury model, an increase in the number of firms reduces the expected benefit to investment (a particular firm is less likely to win), leaving expected costs unchanged. The firm responds by reducing investment. In the Lee-Wilde model, both expected benefits and expected costs are reduced by the addition of another firm (since the flow investment will be made for a stochastically shorter period of time), and the net effect is to enhance incentives to invest. Implications 1 and 1′ are not inherently contradictory; it is quite possible that although the flow rate of investment increases, expected discounted flow costs decrease with an increase in the number of firms. What *are* contradictory are these models' respective implications regarding the effect of an increase in the number of firms upon the expected time until success for an individual firm. Since in both cases, $E(t_i) = 1/h(x_i)$, we see that the fixed cost model implies that $E(t_i)$ increases with the number of firms, while the flow cost model implies that $E(t_i)$ decreases with the number of firms.

When one relaxes the assumption of perfect patent protection, it is easy to construct examples in which an increase in the number of firms decreases the individual rate of investment in a flow cost model; this is because if imitation is a sufficiently attractive alternative, the firm is less concerned about being first (see my 1982 paper). In fact, when imitation is sufficiently swift and complete, there may be an inverse relationship between *aggregate* investment and the number of firms in the industry (Carl Futia, 1980). An alternative form of nonappropriability occurs when rival firms experience significant positive spillovers from each others' research and development expenditures. If these spillovers are sufficiently large, then aggregate investment is inversely related to the number of firms in the industry (Michael Spence, 1982). Thus both the degree of appropriability and the number of firms have direct effects on investment; in addition, there are interaction effects between the degree of appropriability and the number of firms. Since the number of firms engaging in *R&D* is also endogenous, any test of these hypotheses requires a simultaneous equations approach.

2. In a Nash equilibrium with unrestricted entry, there will be excess capacity in the *R&D* technology.

2′. In a Nash equilibrium with unrestricted entry, there will be no excess capacity in the *R&D* technology.

In the lump sum expenditure model of Loury and Dasgupta-Stiglitz, it can be shown that with unrestricted entry, in equilibrium firms will operate their *R&D* projects in a region of increasing returns to scale. In the flow cost model of Lee-Wilde, this result is reversed; firms will always operate in the decreasing returns portion of the innovation production function.

3. At equilibrium, an increase in aggregate rival investment results in a decrease in investment by a single firm.

3′. At equilibrium, an increase in the aggregate rival investment rate results in an increase in the rate of investment by a single firm.

In the fixed cost models, the profit-maximizing investment is smaller the greater is aggregate rival investment, while in the flow cost models, the profit-maximizing rate of investment is greater the greater is the aggregate rival investment rate. Alternatively put,. in the fixed cost models, best response functions are decreasing at equilibrium, while in the flow cost models, they are increasing.

These models all focused upon symmetric equilibria in which no previous innovation was assumed and no future innovation is anticipated. Any stochastic theory of industry evolution will give rise to asymmetric initial conditions; moreover, M. Therese Flaherty (1980) has demonstrated the manner in which industry members who are initially identical may end up following divergent paths even when industry evolution is completely deterministic. In view of these theoretical considerations as well as the obvious empirical fact of asymmetry, it is important to develop asymmetric models of innovation if we wish to apply them to real industries. The models discussed below add such an asymmetry, either through an inherited asymmetric market structure, or through the assumption of a leader/follower, rather than simultaneous-move, framework. The next few implications deal with the impact of current monopoly power and anticipated future innovation upon incentives to invest in *R&D* when firms invest simultaneously.

4. There is an inverse relationship between the magnitude of an innovation and the likelihood that it is invented by a current industry leader.

5. Investment in *R&D* is lower for a large incumbent firm and challengers alike the greater is the flow of current revenue to the incumbent firm.

6. The rate of individual firm investment on a particular innovation declines with the number of anticipated subsequent innovations.

These results follow from the model of a sequence of innovations developed in my forthcoming article. An innovation is termed drastic if the innovator captures a sufficiently large share of the post-innovation market; that is, if the innovation substantially replaces whatever product or process was previously used. When firms anticipate a sequence of drastic innovations, the current industry leader, or incumbent, invests less than each challenger and will thus succeed itself as incumbent (on average) less than $1/n$ percent of the time, where n is the number of firms. The intuition behind this result is straightforward. When invention is uncertain, a firm making higher profits today gains relatively less from invention than a firm with lower current profits; consequently, an industry leader invests less than a challenger or potential entrant. A simple extension of this model indicates that for innovations which are minor (i.e., for which the innovator captures a sufficiently small fraction of the market), the incumbent will invest more than a challenger. Thus we would expect an inverse relationship between the magnitude of the innovation and the likelihood that it is developed by a current industry leader. Moreover, this implication is robust to changes in the specification of costs; that is, this result is insensitive to the fixed versus flow cost assumption. Using a single innovation, fixed cost model with one incumbent monopolist and one challenger, Richard Freeman (1982) found that for large innovations a single challenger will invest more on *R&D* than an incumbent monopolist; Thomas von Ungern-Sternberg (1980) found that for small innovations an incumbent monopolist will invest more than a challenger, and that the probability that the monopolist will succeed first is decreasing with the magnitude of the innovation.

The second implication above is a pure equilibrium effect. An increase in flow revenues to the incumbent has no direct effect upon the challenger's payoff; however, it does induce the incumbent to invest at a lower rate. Because my model (forthcoming) employs a flow cost specification, best response functions are increasing; consequently, the equilibrium response of challengers is to reduce their investment as well.

Finally, implication 6 is a consequence of two effects; a sequence of anticipated future innovations reduces the value of being the incumbent (because no firm can expect a long tenure as incumbent when many innovations remain), and increases the value of being a challenger (because one has many remaining opportunities to succeed). These two effects combine to reduce current investment in *R&D*.

The following implications discuss the impact of changing the timing of the game; suppose that there is a leader/follower structure (in which the incumbent monopolist moves first) rather than a simultaneous-move structure.

7. If the innovation production process is nonstochastic, then a firm which currently dominates an industry will persist as a monopolist, because it will preemptively patent innovations before potential entrants.

8. The above argument holds only if the industry is one in which the threat of antitrust intervention precludes *ex post* negotiation and exclusive licensing. If *ex post* licensing is permitted, the most efficient firm would patent the innovation, but this need not be the incumbent.

The model which generates the first of these implications appears in Dasgupta-Stiglitz (1980), and is more fully developed in Richard Gilbert and David Newbery (1982). Gilbert-Newbery describe a bidding model of *R&D* in which invention is deterministic; thus the question of who invents is essentially one of who has the most to gain from doing so. An entrant will be willing to bid post-innovation duopoly profits for the innovation. By permitting entry, the incumbent and the entrant receive post-innovation duopoly profits; by preemptively patenting the innovation, the incumbent receives post-innovation monopoly profits. Since the present value of monopoly profits exceeds the sum of duopoly profits, the monopolist will bid more for the innovation (i.e., patent it before the entrant). The qualification is voiced by Stephen Salant (1984), who argued that if an entrant anticipates the possibility of innovating and subsequently selling out to the current incumbent, then it will not evaluate the gains from inventing as merely its share of duopoly profits, but will include the expected gains from licensing its patent to the incumbent. In this case, the most efficient firm —not necessarily the incumbent—would patent the innovation.

9. Licensing encourages research when production costs are relatively similar, and discourages research when production costs are relatively disparate.

In a two-firm model of research and development with licensing, Nancy Gallini and Ralph Winter (1983) discuss two incentives to license. The first of these, termed the *ex post* incentive, is the one identified by Salant —the incentive to reduce production inefficiencies and monopolize the output market. There is also an *ex ante* incentive to license (originally identified by Gallini, 1983), which reflects the gain from eliminating wasteful research expenditures as well as the threat of a potentially low-cost competitor. By licensing its technology to a potential challenger at a sufficiently low royalty rate, an incumbent firm can make *R&D* a less attractive prospect to the challenger; this simultaneously reduces expenditures on *R&D*, and removes the threat that the challenger may discover a very low-cost technology. Thus a large *ex post* incentive makes research attractive, while a large *ex ante* incentive reduces the return to research. When production costs are sufficiently similar, *ex ante* incentives are too weak to dominate the gains from *R&D* generated by *ex post* incentives, and investment is encouraged. When production costs are sufficiently dissimilar, *ex ante* incentives dominate, and investment is discouraged. Empirical testing again would require a simultaneous equations approach in which investment in research and development and some measure of licensing behavior are to be explained.

10. Over the course of developing an innovation, the configuration of firms engaging in *R&D* will become more concentrated

as some firms fall sufficiently far behind and, consequently, drop out.

Equilibrium behavior in the models of Drew Fudenberg et al. (1983) and Christopher Harris and John Vickers (1983) is characterized by this pattern. There is an initial burst of investment in which several firms participate; however, when rival firms fall sufficiently far behind the leader, they prefer to drop out of the competition. Consequently, the leader completes the innovation at its preferred, more leisurely, pace. Although extremely stylized, these models incorporate learning and experience in a way not found in previous work.

III. Conclusions

Although individual models have unambiguous implications, the array of existing models still generates considerable controversy. This heightens the interest in and need for empirical tests of these theories. Unfortunately, these implications are generated by highly simplified models, which may make empirical testing more difficult. For instance, some very real aspects of industrial competition are left out, including the possibility of incumbent advantages (for example, better access to capital markets, internal financing, economies of scope) and disadvantages (for example, bureaucratic red tape, weak employee incentives due to a tenuous connection between performance and reward). Also left out are the possible effects of conglomerate diversification; all of these models compare expenditures in a single research area, rather than in the sort of diversified portfolio of projects which might be common among large firms.

Although some sources of data which are suitable for testing these hypotheses do exist, much of the existing data is too aggregated. In addition, many of these hypotheses rely on data which may be difficult to obtain, such as information about the research programs of unsuccessful firms. In order to move in the direction of empirical testing, we must both extend these models in more realistic directions to accomodate existing data, and attempt to gather the specific data required to test directly such models of firm behavior.

REFERENCES

Dasgupta, Partha, "The Theory of Technological Competition," Discussion Paper, International Centre for Economics and Related Disciplines, London School of Economics, 1982.

______ **and Stiglitz, Joseph** "Uncertainty, Industrial Structure and the Speed of *R&D*," *Bell Journal*, Spring 1980, *11*, 1–28.

Flaherty, M. Therese, "Industry Structure and Cost-Reducing Investment," *Econometrica*, July 1980, *48*, 1187–209.

Freeman, Richard, "A Model of International Competition in Research and Development," manuscript, Federal Reserve Board, December 1982.

Friedman, James W., "A Noncooperative Equilibrium for Supergames," *Review of Economic Studies*, January 1971, *28*, 1–12.

Fudenberg, Drew et al., "Preemption, Leapfrogging and Competition in Patent Races," *European Economic Review*, No. 1, 1983, *22*, 3–31.

Futia, Carl A., "Schumpeterian Competition," *Quarterly Journal of Economics*, June 1980, *95*, 675–95.

Gallini, Nancy T., "Strategic Deterrence by Market Sharing: Licensing in Research and Development Markets," manuscript, University of Toronto, 1983.

______ **and Winter, Ralph A.,** "Licensing in the Theory of Innovation," manuscript, University of Toronto, 1983.

Gilbert, Richard J. and Newbery, David M. G., "Preemptive Patenting and the Persistence of Monopoly," *American Economic Review*, June 1982, *72*, 514–26.

Grether, David M., "Market Structure and *R&D*," Caltech Social Science Working Paper No. 58, 1974.

Harris, Christopher and Vickers, John, "Perfect Equilibrium in a Model of a Race," manuscript, Oxford University, February 1983.

Kamien, Morton I., and Schwartz, Nancy L., *Market Structure and Innovation*, Cambridge: Cambridge University Press, 1982.

Kreps, David M. and Wilson, Robert, "Sequential Equilibrium," *Econometrica*, July 1982, *50*, 863–94.

Lee, Tom and Wilde, Louis L., "Market Structure and Innovation: A Reformulation," *Quarterly Journal of Economics*, March 1980, *94*, 429–436.

Loury, Glenn C., "Market Structure and Innovation," *Quarterly Journal of Economics*, August 1979, *93*, 395–410.

Reinganum, Jennifer F., "A Dynamic Game of *R&D*: Patent Protection and Competitive Behavior," *Econometrica*, May 1982, *50*, 671–88.

______, "Innovation and Industry Evolution," *Quarterly Journal of Economics*, forthcoming.

Salant, Stephen W., "Preemptive Patenting and the Persistence of Monopoly: Comment," *American Economic Review*, March 1984, *74*, 247–50.

Scherer, F. M., *Industrial Market Structure and Economic Performance*, 2d. ed., Chicago: Rand McNally, 1980.

Selten, Reinhard, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, No. 1, 1975, *4*, 25–55.

Spence, Michael, "Cost Reduction, Competition and Industry Performance," Harvard Institute of Economic Research, Discussion Paper, 1982.

von Ungern-Sternberg, Thomas, "Current Profits and Investment Behavior," *Bell Journal of Economics*, Autumn 1980, *11*, 745–48.

Field Research on the Link Between Technological Innovation and Growth: Evidence from the International Semiconductor Industry

By M. THERESE FLAHERTY*

During the last decade considerable research effort has been devoted to field studies of the U.S. semiconductor industry and its global competitors. As a result, there is now a large body of data available that should provide new insight into the microeconomics of innovation and growth.

The first step, however, is to provide a parsimonious summary of the innovative process. Empiricists relating firm performance to industry structure and technological effort generally work with published statistics at the industry or firm level. As the samples vary, the estimates have yielded contradictory results. Theorists who struggle with problems of dynamics, uncertainty, and rivalry have a particularly bewildering problem of choice. They are able to work with only a few of many individually plausible and mutually inconsistent assumptions. Field work can help them to focus their efforts on groups of assumptions which are empirically significant. The field workers' conceptual models of how innovation and commercialization occur should be useful to theorists and empiricists. As a result, both will probably be pushed into studying somewhat different problems with somewhat different variables. In addition, empiricists attempting to predict and explain performance should find help in identifying many of the factors which contribute to, but do not alone determine, performance.

In this paper I explore the role field research can play in the effort to understand the microeconomics of technological innovation and growth. Rather than summarizing this burgeoning field, I illustrate the potential of field research by discussing three salient properties of technology and competition in the semiconductor industry which suggest new directions for theoretical and empirical work. I begin with a rudimentary description of the semiconductor industry, its production process, and competitiveness. Then in the sections following, I discuss these three properties.

I. Characteristics of the Semiconductor Industry

In 1979 the total worldwide sales of U.S.-based semiconductor companies were over $8 billion. Industry sales had grown at an average of more than 20 percent each year for twenty years. Value-added per employee of the U.S. merchant component manufacturers had increased over 20 percent, and employment had increased 50 percent between 1975 and 1979. The merchant device manufacturers typically spend more on *R&D* relative to sales than the average U.S. manufacturing company—7 percent rather than 1.9 percent. Semiconductor technology is widely believed to have spurred the development of the computer industry as well as significant advances in consumer electronics and telecommunications. Until the late 1970's, U.S.-based companies dominated the industry technically and commercially. Then, several Japanese companies seriously challenged the U.S. leaders.

The development and manufacture of components entail the design of a component, the fabrication on silicon wafers of the electronic circuit, assembly of the circuit with its connections to the outside in a protective package, and testing of the final product. Fabrication is widely considered to be the

*Harvard University, Graduate School of Business Administration, Soldiers Field, Boston, MA 02163. I acknowledge helpful suggestions by Richard Caves, Anne Coughlan, Annette LaMond, Richard Levin, and Barbara Spencer. My field work in the semiconductor industry was supported by the National Science Foundation Grant PRA-78-21837.

least understood step in the process, the step which defines the state of the art. It is in this step that most of the fabled "experience" in the industry is acquired. And it is in this step that experience in earlier products makes it easier to produce advanced products. The component manufacturers developed most of their own production equipment during the 1950's and early 1960's. But by the 1970's most production equipment and materials were supplied by independent companies, many of them spin-offs from the component manufacturers.

Price-cost margins have been reasonably tight in the industry. In Robert Wilson et al.'s study (1980) over the fourteen-year period from 1964 to 1977, semiconductor manufacturers' average net earnings after taxes (4.5 percent) were less than that for all U.S. manufacturers (5.0 percent). Quality and delivery lead times were critical in some products during the late 1970's. And product life cycles have typically been short relative to those in other industries—five years not being unusual. The cost of building a minimum efficient scale fabrication facility for state-of-the-art devices rose from $1 million in 1970 to over $50 million by the late 1970's. The fixed cost of developing new integrated circuits has also been rising rapidly.

Now I turn to the first of the three properties of the industry which warrant further attention from microeconomists.

II. Appropriability of Technological Breakthroughs

A fundamental problem identified by the microeconomics of innovation is that inventors might not be able to appropriate the returns from their efforts. The patent system which should make appropriability perfect in principle has been government's response to the problem.

But in the semiconductor industry patents have conferred no guaranteed monopoly. John Tilton's study (1971) focussed on the diffusion of semiconductor component manufacturing and marketing capability in the United States, Western Europe, and Japan during the 1950's and 1960's. He found that Bell Labs and the receiving tube companies which had invented the basic products and processes did not succeed commercially relative to the new small firms, for example, Texas Instruments and Fairchild, which developed the inventions further and built new products. Perhaps this was due to Bell Labs' liberal licensing policy. But the large companies entered the product markets late and with less well-adapted products than the new entrants.

From the 1950's on, the very big patentable and scientific breakthroughs did not by themselves confer appropriable competitive advantages on their discoverers. This was partly because many firms were working on closely related projects, so no single firms could get a strong patent position and many were ready and able to copy significant breakthroughs quickly. Of course, being up to date in basic research probably conferred some time and expertise advantage in product development.

This fast-diffusion property of basic research seems to pose a problem for U.S. companies during the 1980's. William Murphy and Joseph Bower's case, for example, illustrates this and related issues from the standpoint of the chief administrative officer of one U.S.-based semiconductor company in 1982. Microeconomists might profitably study the private and social incentives for cooperative efforts at basic and applied research. In this connection the incidence and extent of technological spillovers appears to be critical.

The problem of appropriating innovation appears to have been much less significant with incremental development and engineering innovations—or know-how. The significance and pervasiveness of this phenomenon is reinforced by my 1982 study of ten semiconductor product markets during the 1960's and 1970's. Managers interviewed for the study stated that they derived virtually all of their profits from the sale of their products, not technology licensing, in markets where their know-how distinguished them from their competitors. (See also, for example, the business and technology histories written by Anne Coughlan and myself, 1983, as background for that study.) William Finan's 1975 study of international technology transfer supported this. Similarly, Richard Levin's (1982) historical review of the effect of government actions on semiconductor innovations concluded that patents did not have a significant effect on the development of technology. And Michiyuki Uenohara et al. (1984) found that the Japanese semiconductor industry depended on the large U.S. firms for basic processes and products, while their sales during the 1960's and 1970's turned on product availability, manufacturing know-how, and consumer product designs. Thus it was the slow-diffusing development and know-how aspects of semiconductor technology—not the basic discoveries which diffused quickly—on which companies appropriated returns and grew.

Slow diffusing technology is not only important: it is tractable for further study. Economists have eschewed studying the process of great invention because it seems to have elements of serendipity and peculiarity of technology and "great men." But the process of developing slow-diffusing technological know-how is done by much more comprehensible and predictable engineers. This suggests that microeconomists focus more attention on measuring and understanding these processes and their effects as opposed to patent races. So they might study process engineering and product development rather than aggregate *R&D*. Further, they might consider the effectiveness of increased attention to process improvement during the product life cycles on future products. This also suggests that the simple connections between the effectiveness of basic research, conventional *R&D* resources, and conventional marketing and manufacturing resources may be important in determining the ability of a company to innovate well and to commercialize good products effectively.

III. Interaction Between Technology and Conventional Business Resources

Much of the research on the determinants of the commercial success of technological innovation emphasizes that successful innovations are not only technologically clever, but that they satisfy important market needs. But the nature of this linking may be taken to be simply an appropriate or serendipitous choice of product design (given the available information set) *or* a concerted development, marketing, manufacturing, and distribution effort to make high-quality products in volume and accessible to customers.

The field evidence is enlightening. For example, companies in the semiconductor industry have used their engineering staffs and their favored customers to develop better products. In the equipment-supplying segments of the industry, the extent and nature of the development efforts innovators have made with their users is documented by Eric von Hippel's study (1977). He found that most suppliers worked with device manufacturers to perfect production equipment. A similar process was documented in the silicon wafers segment of the industry in my 1980 study, while Coughlan (1982) documented two such processes for early digital logic circuits. Component manufacturers tended to have far less pre-introduction interaction with customers on product design, but successful companies 1) introduced a number of products while the industry groped for dominance, and 2) worked with customers extensively after the product introduction to aid them in designing new components into their products and troubleshooting problems that arose during their production.

The interaction of technological advantage with conventional business resources was also supported by my study (1983) of market-share determination based on the histories of ten international semiconductor markets. Within each regional product market (i.e., given a dominant design), the length of the technology pioneer's lead showed no positive statistical relation to market share. But the product of the length of the lead and the company's share of applications engineering resources based in the region was positively related to market share. It seems that the portion of long-term loyal customers captured by the innovator during the initial monopoly period depended not on the length of the lead alone, but on the innovators' efforts to aid customers with their engineering problems during it.

My 1983 study also identified several other aspects of technological leadership and conventional business resources as important contributors to market share. Simply being the leader boosted the leader's market share regardless of the length of the lead or the value of other business resources relative to rivals. If one rival had a local marketing subsidiary, then no other firm could penetrate that regional product market unless it had one. Also sales and postsale service efforts (relative to rivals) seemed to boost market share. Finally, Franklin Weinstein et al.'s study (1984) argues that the financial investment and manufacturing quality have also been critical to Japanese achievements.

These results should be viewed as working hypotheses. They suggest that companies develop products and make them commercial successes by using technology in combination with conventional business resources such as marketing, manufacturing and applications engineering in the context of rivals. For microeconomists this should mean more than the relevance of rivalry to the growth and innovation processes. It should also suggest a need to model and measure better firms' conventional business resources and the ways successful companies use them with technology over product lines and over time.

IV. Single Product Innovation and Overall Business Performance

Conventional business resources, so important to combine with technology in the individual product market, can also create significant early mover advantages for two reasons. First, marketing, manufacturing, and applications engineering resources can only be accumulated slowly—and once employees have proven effective, companies are reluctant to let them go for any but long-term reasons. Second, the professionals who comprise these resources are specialized so that within their company they work toward a focussed goal. The focus is an effective way for top management to economize on information processing and attain economies of specialization as well as an impediment to changing the goal.

Coughlan's study (1983) documents how several U.S. semiconductor companies established local marketing subsidiaries with one strong product and then used the facilities to market all their later products. In a following study in 1983, Coughlan and I demonstrated that at least some managers believe that they can create early mover advantages with marketing. For 9 percent of the firms, or 9 out of 102, the choice of whether to establish a sales subsidiary for the foreign region for a particular product was poorly explained by the motives of a firm in a single product market. And, 8 of those 9 established a local marketing subsidiary even though their current product marketing decisions and opportunities did not seem to warrant it. In all these cases the manager of the business had emphasized during interviews his ambitious plans for long-term market penetration.

These findings concur with Tilton's study of the international diffusion of technology. He noted the early presence in Europe of subsidiaries of the U.S. firms. The manufacturing and marketing presence of the U.S. companies may have generated little incremental gain in the markets for products sold at the time of direct foreign investment. But they appear to have established early mover advantages for the U.S. firms which discouraged European potential entrants and made entry difficult for those who tried.

Both focussing and accumulating aspects of conventional business resources are documented by Wilson et al. in their study. They found that one group of merchant semiconductor firms was able continually to generate a sequence of major new products with a particularly heavy investment in *R&D*. Similarly, another group of firms was successful by serving as second sources and low-cost, high-volume suppliers through a strategy of investing relatively less in *R&D* and more heavily in manufacturing capabilities. The firms in the less-profitable, less-innovative segment never moved to the more-innovative group—apparently because their employees as well as their marketing, design, and manufacturing systems were fine-tuned in another, difficult-to-change direction.

These properties should suggest the need for better understanding and measurement of how different long-lived business resources are accumulated and focussed. They raise the question of how a competitive edge could be sustained or eroded—perhaps using dynamic game theory. And they suggest the importance of sequences of products introduced in a process of on-going rivalry—not unlike an industry evolution (see Richard Nelson and Sidney Winter's 1982 approach).

V. The Role of Field Studies in the Microeconomics of Technological Innovation and Growth

Field work should not end with the reporting of what has occurred. Abstracting the important forces from the details involved is a rewarding and creative first step in identifying the problems. In this spirit the abstractions and lessons discussed above are intended to illustrate how field work can focus attention on issues and situations which are important in innovative industrial businesses like semiconductors.

Field work can also identify the important factors which affect performance and competition. This can entail recognizing the severe problems in aggregating applied research and firm-specific development activities in *R&D*, as well as applications engineering and service in marketing. It can also entail recognizing strategic effects and motivations which affect decision making and competition in individual product markets.

Field work can also identify interactions among factors, such as the interaction between the length of technological lead and applications engineering effort in influencing market share. Many of these might be difficult to identify without a detailed understanding of how the industry works. Understanding such interactions through field work is particularly important in cases where the commercial success of a product is due to many factors, no one of which is sufficient.

Of course, making field work's implications accessible to theorists and econometricians has frequently been neglected by field workers, theorists, and econometricians alike. As a result theorists and econometricians have forfeited valuable inputs to their efforts, and field workers have had less impact and leverage on the effort to understand technological innovation and growth.

Finally field work in the semiconductor industry is not a cure-all. The industry is innovative and important, and many of its properties appear to be shared by a large group of manufacturing industries. However significant differences across industries are to be expected. For example, patents appear to be a much more effective means of appropriating returns to innovation in the chemical industries than in semiconductors. Surely, the microeconomics of technological innovation and growth should rely on field studies in the semiconductor and other industries. Feedback between field work and other areas of economic inquiry focussed on technological innovation and growth should prove profitable to all.

REFERENCES

Coughlan, Anne T., "Business and Technology History of Digital Logic Circuits," mimeo., 1982.

______, "International Marketing Channel Choice: The Case of the Semiconductor Industry," Working Paper Series No. AEM 82-08, University of Rochester, January 1983.

______ **and Flaherty, M. Therese,** "Measuring the International Marketing Productivity of U.S. Semiconductor Companies," in David Gautschi, ed., *Productivity and Efficiency in Distribution*, Amsterdam: Elsevier Science Publishing, 1983.

Finan, William F., "The International Transfer of Semiconductor Technology through U.S.-Based Firms," Working Paper No. 118, National Bureau of Economic Research, 1975.

Flaherty, M. Therese, "Business and Technology History of Silicon Wafers for Integrated Circuits," mimeo., 1980.

______, "Market Share, Technology Leadership, and Competition in International Semiconductor Markets," in Richard S. Rosenbloom, ed., *Research on Technological Innovation, Management and Policy*, Vol. 1, Greenwich: JAI Press, 1983.

______, "Determinants of Market Share

in International Semiconductor Markets," in *Race for the New Frontier*, New York: A Touchstone Book, Simon and Schuster, forthcoming 1984.

Levin, Richard C., "The Semiconductor Industry," in Richard R. Nelson, ed., *Government and Technical Progress: A Cross-Industry Analysis*, New York: Pergamon Press, 1982.

Murphy, William J., (under the supervision of Joseph L. Bower) "The Microelectronics and Computer Technology Corporation," Case 0-383-067, Harvard Business School, 1982.

Nelson, Richard R. and Winter, Sidney G., *An Evolutionary Theory of Economic Change*, Cambridge: Harvard University Press, 1982.

Tilton, John E., *International Diffusion of Technology: The Case of Semiconductors*, Washington: The Brookings Institution, 1971.

Uenohara, Michiyuki et al., "Contrasting Patterns of Technological Development," in D. Okimoto et al., eds., *Competitive Edge*, Stanford: Stanford University Press, 1984.

von Hippel, Eric, "The Dominant Role of the User in Semiconductor and Electronic Subassembly Process Innovation," *IEEE Transactions on Engineering Management*, May 1977, EM-*24*, 60–71.

Weinstein, Franklin B., Uenohara, Michiyuki and Linvill, John G., "State-of-the-Art Technology: Strengths and Weaknesses of Each Side" in D. Okimoto et al., eds. *Competitive Edge*, Stanford: Stanford University Press, 1984.

Wilson, Robert W., Ashton, Peter K. and Egan, Thomas P., *Innovation, Competition, and Government Policy in the Seminconductor Industry*, Lexington: Lexington Books, 1980.

Semiconductor Industry Association, "Semiconductor Industry Economic Ratios," Cupertino, CA 1981.

The Relationship Between Federal Contract *R&D* and Company *R&D*

By FRANK R. LICHTENBERG*

According to recent estimates, federal budget outlays for research and development (*R&D*) will increase nearly 19 percent between fiscal years 1983 and 1984. The *R&D* outlays for national defense will account for virtually all of that increase; this component will rise 28 percent, and absorb 70 percent of all federal *R&D* spending in 1984, compared to under 50 percent in 1980. Since work corresponding to about one-half of the dollar value of the federal *R&D* program is generally performed by private industrial firms under contract with federal agencies, the rapid increase in federal *R&D* outlays will presumably be reflected in a similar increase in the value of commitments by industrial firms to perform government *R&D*.

One of the most important economic questions posed by the large prospective increase in resources allocated to federal contract *R&D* concerns its consequences for the economy's rate of technical progress, of productivity growth. In the next section I argue that the ultimate impact on productivity depends on how company decisions to finance *R&D* are affected by the availability of federal contract *R&D* funds. I then provide a brief review and critique of previous studies of the relationship between company- and federal-financed *R&D* performed in industry, and present new estimates which differ from existing ones with respect to both methodology and implications. These estimates are consistent with the hypothesis that increases in federal *R&D* activity tend to be associated with significant reductions in company-financed *R&D*.

I

A number of studies have attempted to determine the *ceteris paribus* effects of company and federal *R&D* on the rate of productivity growth ($\dot{P}$) by estimating variants of the equation

$$(1)\qquad \dot{P} = \alpha_0 + \alpha_1 F + \alpha_2 C,$$

where F and C denote measures of federal and company *R&D* activity, respectively. The principal findings of this literature are that 1) the coefficient on company *R&D*, α_2, is generally positive and strongly significantly different from zero, and 2) the coefficient on contract *R&D* is generally insignificant, sometimes negative, and almost uniformly substantially smaller than α_1; the hypothesis that $\alpha_1 = \alpha_2$ is usually decisively rejected. (There is also evidence that a dollar of company *R&D* expenditure contributes more to the U.S. share in world exports than a dollar of federal expenditure, perhaps as a consequence of their differential impact on productivity.) A hypothesis which could account for these findings (but for which there is little empirical evidence) is that the productivity returns to federal *R&D* are no smaller than those to company *R&D*, but simply more difficult to measure, due perhaps to the wider diffusion of federal returns over time and across industries.

In the absence of any effect of F on C, we could perhaps regard α_1 as capturing the complete or general equilibrium effect of F on $\dot{P}$. But suppose company *R&D* responds to the level of contract research, according to the formula (suppressing for simplicity determinants of C other than F)

$$(2)\qquad C = \beta_0 + \beta_1 F.$$

Then the total derivative of $\dot{P}$ with respect to

*Graduate School of Business, Columbia University, Uris Hall, New York, NY 10027. This research was funded in part through the General Motors Corporation Faculty Research Fellowship, Graduate School of Business, Columbia University, 1983.

F is $d\dot{P}/dF = \alpha_1 + \alpha_2\beta_1$, which depends on the returns to company *R&D* and on the response of C to F, as well as on α_1. In particular, $d\dot{P}/dF \gtreqless 0$ as $\beta_1 \gtreqless -\alpha_1/\alpha_2$. The econometric evidence suggests that α_1/α_2 is quite small and insignificantly different from zero. If it is in fact zero, $\text{sgn}(d\dot{P}/dF) = \text{sgn}(\beta_1)$: federal *R&D* will exert a positive effect on the growth rate of productivity if and only if it tends to stimulate company financing of *R&D*. Stated differently, if the direct social returns to contract *R&D* (measured in terms of its effect on productivity growth) are small relative to the returns on company *R&D*, and if federal *R&D* tends to "crowd out" private investment in research, then increases in federal *R&D* will tend to depress productivity growth. The lower the relative returns, the less crowding out is required for the effect on $\dot{P}$ to be negative. On the other hand, if C is stimulated by F, then the latter will contribute to productivity growth indirectly, despite negligible direct returns.

II

Empirical studies of the effect of federal *R&D* expenditure on company expenditures have been of two principal types. The first, of which the study by Edwin Mansfield and Lorne Switzer (1984) is the sole example, involves simply asking a sample of industrial *R&D* officials how company expenditures would respond to specified, hypothetical changes in federal *R&D* support. On the basis of their survey, these authors concluded that changes in federal support tend to induce changes in the same direction in company spending, although some of their findings seemed to suggest that crowding out may in fact occur. The second, much more popular, approach entails estimation of variants or generalizations of the model (2), using actual data on company and federal *R&D* expenditure, either undeflated or deflated by sales and/or an aggregate price index; these studies may in turn be classified as either aggregate time-series, industry-level cross sectional, or firm-level cross sectional. With the exceptions of Jeffrey Carmichael (1981) and Neil Kay (1979), most analysts have estimated the coefficient on federal *R&D* to be positive and significantly different from zero. In particular, using linear specifications of the relationship between company *R&D* expenditure and contract *R&D* expenditure, David Levy and Nestor Terleckyj (1982) obtained estimates around 0.27 from aggregate time-series regressions, Richard Levin (1980) obtained a similar estimate from industry cross-sectional data, and John Scott (1984) reported estimates around 0.07 from the Federal Trade Commission Line of Business ("below" firm-level) data. These investigators appear to interpret their results as offering at least tentative support for the hypothesis that federal support of *R&D* performed by industry stimulates company *R&D* activity. I have serious reservations about this interpretation, for, as argued below, errors in the specification of the company-federal *R&D* relationship, and/or errors in the measurement of real *R&D* activity by source of financing which these studies may have committed, are likely to impart a substantial upward bias to the estimated coefficients. I focus here on two potential sources of bias—endogeneity of contract *R&D*, and erroneous "deflation" of both types of *R&D* expenditure—and propose appropriate, albeit perhaps partial, remedies, which can be implemented with available data.

Studies of the relationship between federal and company *R&D* are typically predicated on the implicit assumption that the level of contract research may be viewed as exogenous in a model of company *R&D* determination. I submit that the adequacy of this assumption depends critically on the level of aggregation under consideration. At the macro and perhaps to a lesser extent at the industry level, this maintained hypothesis may be acceptable. But the notion that federal *R&D* contracts are distributed among firms in an industry randomly, that is, independently of firm characteristics which determine company *R&D* expenditure, is almost certainly unacceptable. Federal contracts do not descend upon firms like manna from heaven; firms must actively solicit and often compete for contracts.

Suppose firms with given observed characteristics (for example, sales) are heterogeneous with respect to certain unobserved characteristics which influence their "demand" for both company and federal research. Because C and F are usually measured as outlays on "variable" $R\&D$ inputs (labor and materials), it is reasonable to view these variables as being determined within the context of a "short-run" factor demand model, that is, as jointly determined by $R\&D$ input prices and the size of $R\&D$ plant, both of which may vary across firms. Assume further that these unobservables are highly stable over time and therefore can be represented by the set of "fixed effects" $D_1, D_2, \ldots, D_{N-1}$ in the model

$$(3) \qquad C_{it}=\beta_0+\sum_{k=1}^{N-1}\beta_k D_k+\gamma F_{it},$$

where C_{it} and F_{it} denote, respectively, company and federal $R\&D$ performed by firm i $(=1,\ldots,N)$ in year t. If (3) represents the "true" relationship between C and F, estimation of the incomplete model $C_{it}=\beta_0+\gamma F_{it}$ (i.e., ignoring fixed effects) will in general yield biased estimates of γ; I strongly suspect that the bias will be upward. To obtain an unbiased estimate we must be able to observe the entire cross section of firms at least twice—at times t and $t+s$, say. Both Carmichael and Levin had the necessary data but do not report estimates of their models allowing for fixed effects. Scott estimated his model with a kind of "fixed effects," but these were firm effects across its lines of business (LBs) rather than over time; since LBs, as well as firms, may exhibit the relevant type of heterogeneity, that procedure probably attenuated but did not entirely eliminate the bias.

Recall that the objective is to determine the relationship between indices of real input devoted to company and federal research. But due to the absence of "good" $R\&D$ deflators, the indices of company and federal activity typically used are expenditures for $R\&D$, either undeflated or deflated by some price index which is probably a poor proxy for the effective marginal cost of $R\&D$ input, thus introducing error into the measurement of C and F. Clearly, deflating both company and federal expenditures by the same error-ridden deflator induces a spurious positive correlation between measured C and F. If undeflated expenditure data are used instead, the direction of bias depends on the relative elasticities of demand for company and federal $R\&D$. Since the analysis of undeflated or poorly deflated expenditure data is subject to biases of unknown magnitude and direction, it is desirable to supplement it with analysis of direct, albeit partial, quantity indices of $R\&D$ input, such as employment of scientists and engineers. It can also be shown that under reasonable assumptions about the processes determining real input to both activities (for example, sluggishness of $R\&D$ response to changes in sales), the conventional practice of dividing both company and federal $R\&D$ by sales is also likely to produce an upwardly biased estimate of the parameter representing this relationship.

III

This section reports estimates of the relationship between company and federal $R\&D$, based on data compiled by the National Science Foundation. The availability of firm-level panel data on $R\&D$ expenditure, and industry-level panel data on both $R\&D$ expenditure and employment, enables us to investigate the issues of lags, fixed effects, and deflation bias in the analysis of this relationship.

Table 1 presents estimates of pooled regressions of the annual change in measures of industries' company-funded $R\&D$ effort (employment or expenditure) on corresponding changes in the federal-funded $R\&D$. In all regressions, changes in company and federal $R\&D$ are measured relative to total $R\&D$ performance in the previous year, that is,

$$\begin{aligned}\tilde{C} &\equiv (C-C_{-1})/(C_{-1}+F_{-1})\\ &=\Delta C/(C_{-1}+F_{-1})\end{aligned}$$

and

$$\begin{aligned}\tilde{F} &\equiv (F-F_{-1})/(C_{-1}+F_{-1})\\ &=\Delta F/(C_{-1}+F_{-1}),\end{aligned}$$

TABLE 1—REGRESSIONS OF CHANGES IN COMPANY-FUNDED *R&D* EXPENDITURE AND EMPLOYMENT ON CORRESPONDING CHANGES IN FEDERAL *R&D*: NSF INDUSTRY-LEVEL DATA[a]

	Expenditures		Employment	
	(1)	(2)	(1)	(2)
$\tilde{F}$	0.090	0.012	−0.182	−0.306
	(0.90)	(0.13)	(2.22)	(3.57)
$\bar{R}^2$	.1887	4.857	.0885	.3267
$\tilde{F}$	0.059	0.008	−0.198	−0.389
	(0.55)	(0.07)	(1.87)	(3.72)
$\tilde{F}_{-1}$	0.216	0.099	0.204	0.005
	(2.14)	(0.98)	(1.93)	(0.05)
$\tilde{F}_{-2}$	−0.156	−0.190	0.097	0.071
	(1.60)	(1.95)	(0.98)	(0.69)
$\bar{R}^2$	.2875	.5231	.1365	.3931

Note: Col. (1) includes industry dummies; Col. (2) includes both industry and year dummies.

[a] *t*-statistics are shown in parentheses.

where *C* and *F* represent the company- and federal-funded *R&D*, respectively, and subscripts denote lagged values. I consider first regressions of the change in company-funded *R&D* expenditure on the change in federal *R&D* expenditure, both of which, in the absence of a superior alternative, are deflated by the *GNP* deflator. The expenditure regressions were estimated on annual data for twelve manufacturing industries over the period 1963–79. In the first equation, which includes a complete set of industry dummies (thereby allowing for industry differences in average *R&D* expenditure growth over the sample period), the coefficient on federal *R&D* is positive but insignificantly different from zero. The second equation includes also a set of time dummies, which are jointly highly significant; their inclusion results in a substantial reduction in the size and significance of the coefficient on federal *R&D*. Since the restriction that the coefficients on all of the time dummies equal zero is decisively rejected by the data, presumably the second model provides a better basis for assessing the *ceteris paribus* effect of the federal-funded *R&D* expenditure on company-financed *R&D* spending.

Although these estimates suggest that the effect on changes in company *R&D* of contemporaneous changes in federal *R&D* is negligible, one might reasonably hypothesize that company financing responds with a lag (of greater than a year) to changes in contract research expenditures. I investigated this possibility by reestimating the models reported above with two lagged values of the federal *R&D* variable included. As Table 1 indicates, in both models $\tilde{F}$ remains positive but insignificant; $\tilde{F}_{-1}$ is positive in both equations but significant only when the time dummies are excluded; and $\tilde{F}_{-2}$ is in both cases negative and marginally significant. Both models suggest that changes in federal *R&D* financing continue to influence company funding in subsequent years. The sum of the lag coefficients, indicating the effect of a sustained change in federal financing, is, for the restricted model (excluding time dummies), .119, roughly equal to the coefficient on $\tilde{F}$ in equation (1). The sum for the unrestricted model is −.083. Hence the hypothesis that contract *R&D* expenditure stimulates company-funded research appears to be undermined rather than strengthened by allowing for lagged effects.

I consider next estimates of similarly specified models estimated using data on employment of *R&D* scientists and engineers, rather than *R&D* expenditure, by source of funding. As noted above, due to the difficulty of properly deflating *R&D* expenditures, these employment data may serve as a better index of real *R&D* input than inaccurately deflated cost series. Table 1 also indicates that when $\tilde{C}$ is regressed on industry dummies and $\tilde{F}$ only, the coefficient on the latter is negative and significant at the 3 percent level; when the time dummies are added, the coefficient increases 60 percent in absolute value. These estimates thus obviously imply a strong, negative contemporaneous effect of federal-funded on company-supported employment. When lagged values of $\tilde{F}$ are introduced, the restricted model produces a coefficient on $\tilde{F}_{-1}$ opposite in sign and roughly equal in magnitude and significance to the coefficient on $\tilde{F}$; this model implies essentially no, or perhaps a weakly positive, net effect of federal *R&D*. The unrestricted model, however, points to a rather different conclusion: although positive, both lag coefficients are far from significant, and the sum of the coefficients, $-.313$, is close to the coefficient on $\tilde{F}$ in the equation excluding lagged values. This regression implies that a federally funded increase of 100 *R&D* scientists and engineers this year will result in a reduction of company-sponsored employment of 39 within the year, essentially no change next year, and an increase of 7 in two years.

The third and last type of data analyzed is longitudinal, firm-level data on *R&D* expenditures as a fraction of sales, obtained from NSF survey responses; these data are described in Zvi Griliches and Bronwyn Hall (1982). Because the data are available only in moment-matrix form to avoid disclosure, my ability to experiment with alternative functional forms was limited; I report here regressions of the ratio of company *R&D* to sales on the ratio of federal *R&D* to sales, for the years 1967, 1972, and 1977, and corresponding regressions involving *changes* in these variables during the periods spanned by these years. Estimating models in both level and first-difference form reveals the importance of allowing for fixed effects in the analysis of *R&D* performance.

TABLE 2—REGRESSION OF COMPANY *R&D*/SALES ON FEDERAL *R&D*/SALES[a]

Year	Federal *R&D* / Sales	Constant
1967	0.046	0.030
	(2.11)	(11.51)
1972	0.100	0.026
	(4.72)	(11.40)
1977	−0.218	0.032
	(5.03)	(11.70)

[a] NSF firm-level data ($n = 991$); t-statistics are shown in parentheses.

TABLE 3—REGRESSION OF CHANGE IN COMPANY *R&D*/SALES ON CHANGE IN FEDERAL *R&D*/SALES[a]

Period	$\Delta\left(\frac{\text{Fed. } R\&D}{\text{Sales}}\right)$	Constant
1967–72	−0.476	−0.001
	(21.84)	(2.26)
1972–77	−0.168	−0.000
	(14.05)	(0.75)
1967–77	−0.261	−0.001
	(14.06)	(2.26)

[a] See Table 2.

Estimates of the level version of the model are reported in Table 2. (All of the equations shown in this table were also estimated including other regressors, such as total employment or sales; the coefficients on the federal *R&D* variable were always essentially unchanged by their inclusion.) In the 1967 and 1972 cross-sectional regressions, the coefficient on federal *R&D*/sales is positive and significant, indicating that firms which do more contract *R&D* tend to perform more own-financed *R&D*. It is rather surprising that the coefficient in the 1977 equation is negative, as well as larger in magnitude and more significant than the 1967 and 1972 coefficients. Only Carmichael has found negative and significant cross-sectional coefficients on contract *R&D*, and his estimates are smaller (of the order $-.08$).

Estimates of the first-difference form of the model for 1967–72, 1972–77, and 1967–77 are presented in Table 3. The coefficients on federal *R&D* in all three equations are negative and highly significant. The point estimate of $-.26$ for the ten-year interval as a whole is similar in magnitude to the coefficients in the *R&D* employment regressions reported above.

IV

The failure of most econometric studies to find significant direct productivity effects of federal contract *R&D* has led some analysts to hypothesize and investigate an indirect path of stimulus to productivity, via an inducement to perform company *R&D*. The evidence presented in this paper—based, I believe, on improved measurement of, and specification of the relationship between, company and federal *R&D*—is markedly inconsistent with this hypothesis. These findings thus make heavier the burden of proof on those who would claim that federal contract *R&D* makes a positive contribution to aggregate technical progress.

REFERENCES

Carmichael, Jeffrey, "The Effects of Mission-Oriented Public *R&D* Spending on Private Industry," *Journal of Finance*, June 1981, *36*, 617–27.

Griliches, Zvi and Hall, Bronwyn, "Census-NSF *R&D* Data Match Project: A Progress Report," paper presented at the Census Workshop on the Development and Use of Longitudinal Establishment Data, January 1982.

Kay, Neil M., *The Innovative Firm, A Behavioral Theory of Corporate R&D*, New York: St. Martin's Press, 1979.

Levin, Richard C., "Toward an Empirical Model of Schumpeterian Competition," unpublished paper, Yale University, May 1980.

Levy, David M. and Terleckyj, Nestor E., "Effects of Government *R&D* on Private *R&D* Investment and Productivity: A Macroeconomic Analysis," paper presented at annual meeting of Southern Economic Association, November 1982.

Mansfield, Edwin and Switzer, Lorne, "Effects of Federal Support on Company-Financed *R&D*: The Case of Energy," *Journal of Management Science*, forthcoming 1984.

Scott, John, "Firm versus Industry Variability on *R&D* Intensity," in Zvi Griliches, ed., *R&D, Patents and Productivity*, Chicago: University of Chicago Press, forthcoming 1984.

PSYCHOLOGICAL AND SOCIOLOGICAL FOUNDATIONS OF ECONOMIC BEHAVIOR

Gift Exchange and Efficiency-Wage Theory: Four Views

By GEORGE A. AKERLOF*

My earlier paper (1982) viewed the labor contract as a "partial gift exchange." According to this view some firms willingly pay workers in excess of the market-clearing wage; in return they expect workers to supply more effort than they would if equivalent jobs could be readily obtained (as is the case if wages are just at market clearing). This partial gift exchange hypothesis is one of several efficiency-wage theories which explain why wages exceed market clearing, or alternatively stated, why there is involuntary unemployment. This paper gives some further commentary on partial gift exchange. A complementary paper in this issue (Janet Yellen, 1984) reviews these efficiency-wage models in general.

If there is involuntary unemployment in an equilibrium situation, it must be that firms, for some reason or other, wish to pay more than the market-clearing wage. And that is the heart of any efficiency-wage theory. Yet there is a natural reason why economists have hitherto resisted the application of efficiency-wage theories to unemployment in developed countries. A view that any buyer should willingly pay more than necessary to any seller seems highly counterintuitive in the paradigm of standard economics, which is that of supply and demand. The purpose of this paper is to show that in the context of four other paradigms of the labor market, such wage-setting behavior is in fact natural. Furthermore, the literature in each of these paradigms marshalls considerable empirical evidence supporting the hypothesis that some firms may pay more than market-clearing wages.

The four paradigms to which I refer are those of dual labor markets, the theory of bureaucracy, the theory of work groups, and equity theory.

I. Dual Labor Market Hypothesis: The First Paradigm

Because of its familiarity to economists, my comments on the dual labor market hypothesis of Doeringer and Piore will be brief. According to this theory there are two types of jobs—those in the primary sector and those in the secondary sector. Primary sector jobs have stability, low quit rates, good working conditions, promotions according to a promotion ladder, acquisition of skills, and good pay. In contrast, secondary sector jobs have high quit rates, harsh discipline, little chance of promotion, low acquisition of skills, and poor pay. The difference between good pay and poor pay between primary and secondary sector jobs can be seen as the difference between wages in excess of market clearing and wages at market clearing. Provided primary sector firms set the wages they "prefer," the dual labor market hypothesis is itself an efficiency-wage theory of the labor market. If the dual labor market hypothesis is not counterintuitive, then neither are efficiency-wage theories of unemployment. Furthermore, all the empirical studies supporting the dual labor market hypothesis also *ipso facto* support efficiency wage theories of unemployment—since primary sectoı firms are paying wages in excess of market clearing.

*Department of Economics, University of California, Berkeley, 94720. I thank Janet Yellen for valuable comments and the National Science Foundation for invaluable financial support under Research Grant No. SES 81-19150 administered by the Institute of Business and Economic Research of the University of California-Berkeley.

I should also mention that the theory of internal labor markets follows closely the basis for the sociological theory of organizations: the primary sector organization described by Doeringer and Piore fits closely the classic description of the bureaucratic organization by Weber.

II. Weberian Theory of Organization: The Second Paradigm

The Doeringer-Piore model is based, in most respects, on Weber's description of bureaucracies. According to Weber, a bureaucracy is a hierarchical organization in which officials follow career paths according to the organization's promotion ladder. There is a well-specified division of labor; and the officials of those organizations exercise an impersonal discipline in the discretionary conduct of their own offices as well as in the exercise of commands from higher offices. This impersonal discipline is the by-product of the personal loyalty of the employees to the goals of the bureaucracy. The most essential feature of the gift exchange model is the importance of employee loyalty to the operation of the firm. Thus the abundant evidence which shows the accuracy of Weber's description of bureaucratic organizations serves as evidence for the empirical importance of this essential feature of the gift exchange model as well.

In recent years empirical sociologists have, however, modified the Weberian theory of bureaucracy through their studies of work groups. These studies (in contrast to Weber's emphasis on hierarchical control) suggest that in the typical organization superiors have only limited control over the work activity of their subordinates. The Doeringer-Piore description of internal labor markets also incorporates this modification due to the study of work groups.

III. Work Groups: The Third Paradigm

Detailed sociological studies (beginning with the examination of the Hawthorne works by Roethlisberger and Dickson) have shown considerable discrepancy between *formal* and *actual* authority in many different work situations. In the work situations examined, workers have, with varying degrees of openness, set their own informal work rules which are often at variance with the official work rules. The ability of management to make workers conform to their authority is far from complete; instead, most studies have shown a complex equilibrium in which official work rules are partially enforced, existing side by side with a set of customs in the work place which are at partial variance with the work rules, and some individual deviance from both the official work rules and the informal work norms. This incompleteness of authority in the work place should not be a surprise. The occasional spectacular jail breaks from even the most closely guarded prisons suggest that authority over subordinates in even the most total institutions is less than complete.

Within this framework where adherence to authority is by nature less than complete, the loyalty of employees is one contributor to high productivity. According to the basic idea of the "labor market as partial gift exchange," the loyalty of workers is exchanged for high wages, and this loyalty can be translated via effective management into high productivity. This is an abstract concept which may be more convincing if it is discussed in the context of a real work situation.

A recent sociological study by Michael Burawoy (1979) repeats a classic study by Donald Roy (1952) of the operation of a piecework machine shop. (At the time of Burawoy's restudy, this machine shop produced parts for truck engines.) The workers in this shop were rewarded for their efforts by an incentive system. According to this payment system, each job was rated for normal production. Workers who produced less than this normal production were awarded their base pay, and workers who produced in excess of this base were awarded proportionately higher wages. Workers used this incentive system to relieve the boredom of their jobs by turning their work into a game. Those who found that they could produce more than the base production engaged in "making out at the game." This making out at the game involved attempting to produce as much as 140 percent of normal production. (Production beyond this level on any day was not recorded because of fear that the rates would be consequently revised.)

As described by Burawoy, this game was played intensely. Workers' conversations at lunch were dominated by discussions of their difficulties or successes in making out. The social exchange of the machine shop was dominated by it. Operators played the game, while various auxiliary workers played a role in either helping or hindering the operators at their making out—playing a role not unlike that of the Chance and Community Chest cards in Monopoly. I will describe these interactions with auxiliaries in a bit of detail, because they indicate the character of the game. These interactions indicate as well that conflict in the work place may as naturally involve workers on the same hierarchical level as superiors and subordinates.

Before starting a new job, typically at the beginning of a working day, an operator had to be assigned by the scheduling man, whose discretion in making easy or difficult assignments had considerable influence on the operator's subsequent success at making out. Before beginning the job he then, typically, had to engage the aid of the crib attendant, whose job included handing out blueprints and tools, some of which might be scattered about the shop floor; the truckers, whose job was to bring stock from the aisles; the setup man, whose job was to help workers set up their machines; and the inspector, whose approval of the first piece was necessary before the operator could engage in subsequent production. In each case the auxiliary worker had considerable discretion as to which worker he might be helping at a given time, and the amount of aid he was going to give. Failure to engage the crib attendant, for example, in grinding tools might cause considerable delay, as could failure to engage a trucker to bring the raw material. The setup man could help the operator make a rapid start. Inspectors could use discretion to accept or reject pieces which were at the margin of the blueprint specifications. In the cases of the scheduling man and the inspector, workers had direct methods of retaliation for unfair treatment: low production would reflect badly on the scheduling man who was responsible for the shop's production, and workers could retaliate against too severe inspection by turning out scrap after the first piece was OK'd, which would look bad for the inspector. Workers did not have direct implicit threats, however, which would enforce equitable treatment from the crib attendant, truckers, or setup men.

Having described this game and the context in which it is played, the question arises whether the concept of wage-induced loyalty can play a role in this machine shop. The operators play the game, which "eliminates much of the drudgery and boredom of industrial work" (Burawoy, p. 89). Workers have converted their job into a type of pinball, with their eye on the score—only the job is dirtier, heavier, more dangerous and more enduring, and the output has value for other persons. Burawoy (p. 89) also assures the reader that workers' desire to make out (i.e., to reach 140) is not due to the added monetary rewards, but rather due to the same type of pleasure as comes from breaking a record in pinball. In such an environment, why would it pay a firm to give more than the minimum wage necessary to attract workers to the plant?

As we have seen, Burawoy's workers have considerable freedom in the operation of their machines and in their complex interactions with other workers. If they choose to produce less, they can do so; in fact Burawoy has described in detail instances in which workers broke the administrative rules to increase their own output (and make out at the game). Thus workers would have no difficulty in *decreasing* their output if they so desired while still abiding by the rules. But why should the wage have any effect on this desire?

If wages are sufficiently low, workers will feel unfairly treated. Such unfair treatment will take the fun out of playing a game whose results benefit the firm. This reduction in fun will have the immediate effect of less willingness by workers to make out. As Burawoy writes a bit abstractly: "The day-to-day experience [of making out at the game] emerges out of the organization of work and defines the interests of the various agents of production *once their basic survival—which, as far as workers are concerned is an acceptable wage—is assured*" (p. 85, emphasis added). Given that the term "acceptable" is a fuzzy concept, so that higher wages have higher probability of being seen as acceptable, this suggests that higher wages will result in increased productivity. Nor should it be forgotten that employees who feel they are unfairly treated will not only fail to indulge in the game, but may also actively participate in changing the rules so that its outcome is less advantageous to the firm.

This picture of gaming and equity yields a more sophisticated version of gift exchange than in my earlier article. It also yields a more sophisticated view of the bureaucratic firm's reaction to incomplete control over work than in the economic articles on shirking. According to this view, in the case where workers have animosity toward their employer, higher wages will cause workers to feel less badly about relieving their boredom by playing a game which yields a surplus to the firm. Or, alternatively, if workers have loyalty to their employer, low wages will cause workers to feel less badly about playing a game which fails to benefit the firm.

In either case, the model in which high wages legitimize the workers' positive feelings for outcomes which benefit the firm can be summarized by an individual utility function

$$U=U(\omega,e;\overline{\omega},u). \tag{1}$$

The utility of the worker depends on his real wage ω, his effort e, the wage paid to other workers $\overline{\omega}$, and the unemployment rate u. The worker at a given firm who maximizes this function will let his effort expended be a function of the real wage paid, the wage others are paid, and the unemployment rate

$$e=e(\omega;\overline{\omega},u). \tag{2}$$

This last equation (2) is, of course, the key ingredient in an efficiency wage model of unemployment.

The gaming in the Roy-Burawoy studies is special. But a general point emerges from this situation which applies to all jobs in which the worker has some degree of freedom. Workers in such jobs can use this freedom to make their workday more pleasant. In most jobs keeping busy makes the time go faster. (Psychological experiments show that subjects who are kept busy estimate a shorter elapse of calendar time than subjects who are idle.) Payment of a fair wage legitimizes for the worker the use of this busyness for the advantage of the firm. Not only may workers keep themselves busy in technical operations (as the operators at the machine shop) but they may be socially busy. For example, George Homans' "Cash Posters" (1954) describes a group of clerical workers who spend their time working quickly, and also engaging in considerable social interaction in the process.

IV. Equity Theory: The Fourth Paradigm

The discussion of wages as contributing to job satisfaction which enhances worker willingness to engage in productive busyness on the job brings up an empirical question. *All things being equal*, do workers with greater pay produce greater output? Social psychologists in equity theory have conducted experiments to establish the empirical validity of such a connection. The classic study is by J. Stacy Adams (1965). (For a recent review of this literature, see Richard Mowday, 1979.)

Adams conducted an experiment in which students were hired for proofreading. One group was told that they were not qualified, but would be paid the usual rate. Another group was told that they were qualified and were also paid the usual rate. Those who were led to believe they were overpaid produced fewer errors when paid on a piece rate basis, and more output per hour when paid on an hourly basis than those who were told they were qualified and received the market rate. Many variants of this experiment have been conducted, some aimed at removing the reduction in self-esteem caused by telling some students they were not well-qualified. Not all of these studies reproduce the result that "overpaid" workers will produce more, but, as might be expected, the evidence appears strongest for the withdrawal of services by workers who are led to believe they are underpaid.

V. Conclusion

Although to an economist the payment to a seller of factor services of more than the market-clearing price seems counterintuitive, at least four paradigms suggest either the empirical existence of such payments or why they will frequently occur in the labor market. In these paradigms efficiency-wage theories of unemployment are natural.

Finally, lest it be thought that these theories involve only *real* variables and thus only describe a *natural* rate of involuntary unemployment, I should add that in such models behavior which will result in small losses to agents will allow large changes in unemployment due to changes in real demand. And thus demand-generated cycles will result if there is near-rationality. (See Yellen, and my paper with Yellen, 1983.)

REFERENCES

Adams, J. Stacy, "Inequity in Social Exchange," in L. Berkowitz, ed., *Advances in Experimental Social Psychology*, Vol. 2, New York: Academic Press, 1965, 267–99.

Akerlof, George A., "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics*, November 1982, *97*, 543–69.

______ **and Yellen, Janet L.,** "The Macroeconomic Consequences of Near-Rational Rule-of-Thumb Behavior," mimeo., University of California, 1983.

Burawoy, Michael, *Manufacturing Consent*, Chicago: University of Chicago Press, 1979.

Homans, George C., "The Cash Posters," *American Sociological Review*, December 1954, *19*, 724–33.

Mowday, Richard T., "Equity Theory Predictions of Behavior in Organizations," in Richard M. Steers and Lyman W. Porter, eds., *Motivation and Work Behavior*, New York: McGraw-Hill, 1979, 124–46.

Roy, Donald, "Restriction of Output in a Piecework Machine Shop," unpublished doctoral dissertation, University of Chicago, 1952.

Yellen, Janet L., "Efficiency Wage Models of Unemployment." *American Economic Review Proceedings*, May 1984, *74*, 200–05.

Introducing Social Structure into Economic Analysis

By JAMES S. COLEMAN*

What I want to do in this paper is to expose some of the social assumptions on which economic analysis depends, first to suggest that these assumptions have allowed economics to make important strides in social theory, but also to suggest that further progress lies in modifying or discarding those assumptions.

Perhaps the best way to begin is to describe a set of episodes or events. Some of these may seem to have little to do with economics, but I ask your patience; I believe the relevance will become clear shortly.

I. The Social Organization of Trust

Episode 1: One December afternoon in 1903 a musical extravaganza was playing before a packed audience at the Iroquois Theater in Chicago. A fire broke out in the stage draperies, someone in the audience yelled "Fire!" and the audience began to panic. The comedian Eddie Foy was on stage, and he attempted to quiet the crowd. He failed, the crowd panicked, and before the panic was over, 587 people had died—most not killed by the fire itself, which was put out shortly, but in the process of trying to escape.

Episode 2: In 1717, a Scotchman named John Law got the Regent in France to charter the Mississippi Company for exploitation of the Mississippi Territory. There was an extraordinary growth of stock speculation, with about 500 stock-jobbers setting up stalls in the gardens of the Hotel de Soissons in Paris, and all of Paris society entrusting their fortunes to John Law and his Mississippi scheme. The trust placed in him was so great that according to one report, he became the most influential person in France at the zenith (see Charles Mackay, 1892).

Then the bubble broke. There were rumors of the failure of the scheme, and all rushed to sell their securities. The panic was so great that stations in life were forgotten; great ladies and footmen alike rushed desperately to unload their shares before the prices plummeted further and they were ruined.

Episode 3: In the late eighteenth century (a century of great innovation in financial institutions), Bank of England notes circulated throughout England, as did notes issued by local banks. In addition, bills of exchange circulated as payments for debts in certain areas of dense manufacturing industry, such as Lancashire.

In English towns some distance from London, Bank of England notes were accepted only at a discount of about 15 percent relative to the notes issued by local merchant banks or goldsmiths. Similarly, in Lancashire, the notes of local merchant banks were accepted by manufacturers only at a discount of about 15 percent relative to the bills of exchange, which were obligations of manufacturers, endorsed by all those manufacturers through whose hands the bill had passed (T. S. Ashton, 1945). As these discounts indicate, there were extensive variations in the confidence placed in the ability of different institutions to repay their obligations.

Episode 4: In many villages in Japan and Southeast Asia, there are "rotating credit associations," which are semi-social and semi-economic institutions that operate as follows: a group of ten to twenty neighbors arranges to meet once a month for a social occasion at one member's home, an occasion at which they not only enjoy food and each other's company, but also each puts a small amount of money into a pot. One "winner" is chosen, usually by lot, to receive the total amount. In subsequent meetings, his name is excluded from the drawing, so that ultimately, all members win one time (John Embree, 1939). This is an important institution for borrowing and for capital accumulation among groups in which

*Department of Sociology, University of Chicago, 1126 East 59th Street, Chicago, IL 60637.

banks or other formal financial institutions are not able to subsist. It allows small capital expenditures (such as for a bicycle), which would otherwise require a greater capacity to save than exists in such semi-subsistence situations.

The dependence of such associations on a very extensive allocation of trust is obvious. Because an early winner could abscond with his winnings, the instruction requires trust of each in each other, and trustworthiness of each. The consequences are equally obvious: the poor in many areas where such trust does not exist (as, for example, in the cities of many countries) are deprived of a valuable economic resource.

What is common to all these episodes is something we could call confidence or trust. In Episodes 1 and 2, escape from a fire and escape from a collapse of a company's stock, there was a sudden withdrawal of trust, as the basis on which trust had been founded was suddenly removed. Yet the episodes concern quite different entities in which trust was placed: a building or an economic enterprise. Episodes 3 and 4 show variations in the stable social organization of trust: variations in eighteenth century England in confidence in the notes, or promises to pay of various institutions, and variations in the twentieth century in the amount and distribution of trust in certain parts of the underdeveloped world.

Given that both the functioning of economic institutions and the theory of such functioning assume a foundation of trust, the further question arises: is economic analysis able to deal with behavior, such as that described in these episodes, in which this foundation can no longer be assumed? I think the answer is both Yes and No. Economic analysis is able to deal with *individual* behavior based on incomplete trust or confidence. The large body of work on decision making under uncertainty or risk is directly applicable. In addition, some recent work, such as that on the principal-agent problem, concerns the optimum use of incentives by a principal to induce trustworthiness on the part of an agent. But economic analysis has not been able to cope with the social organization of trust. One, but not the only, consequence of this is an inability to deal with dynamics as exhibited in the case of panic.

The principal means by which economic theory moves from the micro level of a single economic actor to the macro level involving many such actors is through the ubiquitous concept of a "representative agent." Yet simple aggregation is clearly inappropriate for phenomena such as trust, since trust is a *relation* between two actors. Even more: one actor's placement of trust in a second may be conditional upon that of a third. As a consequence, withdrawal of trust by one actor in a system may have a domino effect throughout the system. It all depends, not simply on the average *level* of trust, but on the *social organization* of trust.

As a result of not explicitly incorporating this social organization into economic theory, there remain many problems, important to economics, that cannot be treated by economic theory. For example, consider Lancashire in the eighteenth century. The acceptance of bills of exchange was based on chains of trust that followed chains of production, while acceptance of the note of a bank is based on trust in a single institution. What is the structure of trust placement under which one kind of obligation will be more acceptable than the other? And what are the consequences of there two patterns of obligations for stability of trust? As in electrical grids or networks which span a number of cities, certain configurations are highly sensitive to breakdowns at a single point, others are not.

II. Markets with Structure

The intellectual problem involved here is a much more general one. I use the examples of trust only because they exemplify it well. The problem is this: we understand and can model behavior at the level of individuals, but are seldom able to make an appropriate transition from there to the behavior of the *system* composed of those same individuals.

Yet the principal intellectual feat of neoclassical economic theory was to do precisely this, with the Walrasian model of an exchange economy. It did so, of course, in an idealized social system: one in which actors were independent; goods being exchanged were private, and tastes were fixed. In this idealized system, it was found to be possible to begin with a distribution of goods among actors, and to end with a set of equilibrium prices and an equilibrium distribution of goods.

Close inspection of this theory can give some indication of the extent of the social assumptions. There are, it is assumed, no social barriers to inhibit information flow and exchange agreements; there is complete intermixing among a large set of independent actors; there are no consumption externalities, that is, no social interdependencies in consumption; the goods exchanged are alienable, and not inherently attached to the person, as is true for labor services; and others.

These assumptions have served economics well, providing a powerful engine for making the micro-to-macro transition. But it is precisely the maintenance of these assumptions that restricts the predictive power of economics, putting certain economic behavior out of its reach.

For example, the theory assumes no interdependencies affecting the exchanges. Yet we all know that persons are resources for one another, and a given person values certain persons more than others. An illustration of the economic consequences of this can be seen in the housing market in a population with two or more ethnocentric ethnic groups. The geographic distribution of households typically shows a high degree of ethnic concentration in neighborhoods, with considerable stability over time. But this stability is punctuated by periods of rapid housing turnover, when one ethnic or racial group, expanding in size, succeeds another.

It is not difficult to incorporate this valuation of "geographically proximate others" into a model of rational action. What is necessary, however, is to incorporate this into the functioning of a *system* of behavior —a housing market. Thomas Schelling (1978) has developed a simulation for such markets to show the degree of segregation that can be generated by even a small amount of ethnocentrism. This is a first step, but a long way from a theory of such markets. The difficulties lie, of course, in the fact that actions are no longer independent as assumed in the neoclassical perfect markets. If the appropriate theory for such markets with interdependencies was in existence, it would have usefulness not only for residential neighboring, but for industrial neighboring as well.

Another example of markets in which social structure is important is the matching process that occurs in monogamous marriage or in job markets. Sociologists and demographers call this assortative mating. It is clearly a social process of some complexity. It can, however, be seen as an economic exchange market in which each actor has exactly one good to trade, and can get exactly one in return. Yet it is very different from a neoclassical perfect market. For example, the role of "price" as an allocation mechanism is greatly altered; and the entities exchanged are not fungible—there is not a market in trading of wives.

Here is an area in which work has been done, both by demographers and by economists. The problems are not solved, but enough has been done to see both the difficulties that arise in constructing a theory of assortative mating markets and some of the benefits of doing so. For example, such a theory in labor markets would help resolve the disagreement between the belief in structural sources of unemployment and unemployment as due to deficient demand. So-called "dual labor market theory" would be replaced by the ability to describe the actual degree and character of segmentation in the labor market.

What I have been describing is the problem of moving from a model of individual behavior to a theory of the behavior of a system composed of these individuals, taking social organization explicitly into account in making this transition, rather than assuming it away. This, I believe, is the central intellectual problem in the social sciences. But it is too often dealt with by fiat, as economists do when they invoke a representative agent to get from a micro level to a macro level. And it is too often ignored altogether, as quantitative sociologists do when they concentrate wholly on explaining individual behavior.

III. Firms as Collectivities

The micro-to-macro problem can be seen in another context, which sociologists call organization theory and economists call the theory of the firm. In the neoclassical theory of the firm, the relative amounts of each factor input demanded by the firm depend on the marginal productivities and their prices, and the quantity produced depends upon the marginal price of the product relative to marginal costs of the factor inputs. In short, the firm is a single actor maximizing its utility—called profit in the case of the firm. In sociology also, most "organization theory" is really decision theory of managers. Here is a second implicit model of social organization to allow moving from the behavior of individuals—that is, the individuals who comprise the firm—to the behavior of a system, which in this case is called a firm. The assumption is the opposite of that used for the perfect market: the "perfect firm" is organized as if it were a single authoritative individual: a complete hierarchy which can be modelled as a single rational actor. The household, even though it is a set of individuals with differing interests, is similarly treated by economic theory, as a single actor.

Yet those social organizations that we call firms or bureaucracies never were such simple entities, and are decreasingly so. In firms, apart from problems of separation of ownership and control that arise in publicly owned corporations, there are developments such as the introduction of some aspects of a market into the firm's structure, through profit centers and the use of transfer pricing. And there are questions of the distribution of resources available to different "stakeholders" in a firm. This resource distribution changes with legislation like the Wagner Act in the United States in 1936, or the 1976 codetermination law in Germany. The codetermination law explicitly introduces a formally democratic system into firm decision making, with workers and stockholders each having representation, and an allocation for votes. In households, the assumption of a single authoritative decision maker has been far from reality for a long time.

Some theoretical work has been done which abandons the single authoritative actor assumption. Part of this work had its origins elsewhere in economics. Kenneth Arrow, examining the properties of a social welfare function which would translate individual preferences into a social choice, showed that there was *no* choice procedure other than the single authoritative actor which obeyed minimum conditions of rationality. The question remains, how to introduce into the theory of the firm the social processes which generate firm behavior when there is, for example, formal democracy created by codetermination laws, or the internal market mechanisms of profit centers with transfer pricing.

There are organization theorists, both in economics and in the other social sciences (for example, Michel Crozier, Oliver Williamson, William Niskanan, Peter Blau) who have taken steps in this direction. But as with the introduction of social organization into the theory of markets, the introduction of markets and other social processes into the theory of the firm is hardly central to work in economic theory. For the behavior of the household as something other than a single actor, work has also been done, for example, Gary Becker's work on the family.

One might say, of course, why make such attempts since the traditional assumptions about social organization have served economists well? Perhaps this was so, when the discipline was in its infancy; but I believe the failure to modify these assumptions constitutes a serious impediment to the policy usefulness of economic theory. Let me mention an example that involves the phenomena I have discussed: firms, markets, and trust.

A major reason for backward vertical integration of firms, incorporating suppliers within the hierarchical organization, is to be able to exercise greater administrative control of scheduling, quality, and meeting of design specifications. The arguments for backward integration have to do with transaction costs, which include these uncertainties and unpredictabilities involved in dealing with independent firms. But integration is done at the cost of sacrificing the economic benefits of a market, which prevents monopolistic behavior on the part of a supplier. Once a productive activity is internalized within a firm, it has a partial or complete monopoly vis-à-vis the departments it supplies. Even with decentralization of the organization, great difficulties arise in establishing appropriate transfer prices in the absence of a true market, and in the presence of the interests of each department in setting as high a price on its services as possible.

Many of the benefits of a hierarchical organization without the disadvantages can be achieved if there is a high level of trustworthiness (in the sense of meeting design, scheduling, and quality obligations) on the part of independent organizations that could supply parts and services. This exists in Japan to a much greater extent than in the West and has been noted by various analysts of Japanese industry (see Rodney Clark, 1979). The result is that Japanese automobile companies (to use one industry in which Japan has been particularly successful) are *not* vertically integrated, as are those in the West. They are principally automobile assemblers, buying most of their parts from independent supplier organizations whose prices are disciplined by the market.

The problem that this example illustrates for the usefulness of economic theory is that the theory has no way, for example, of taking observations on relations of trust and trustworthiness between suppliers and customers in a young industry and predicting the equilibrium degree of vertical integration in the industry.

IV. Conclusion

What I have tried to do in all this is to indicate something about how social organization can be most profitably incorporated into economic theory. This is not by abandoning the conception of rational action of individuals, but by changing the organizational assumptions that translate individual action into systemic or collective action. In doing this, I do not want to obscure one point: the major contribution to theory in social science made by economics through the creation of a micro-to-macro engine, a market conceived as a fully communicating set of independent actors with selfish tastes and private goods. In other areas, such as aggregation of preferences concerning a social choice, and contribution of independent individuals to a public good, economists have shown that the micro-to-macro engine as currently conceived will not work. In still other systems, such as those involving placement of trust, they have not started. Thus the overall task is hardly begun.[1]

Economists have a branch of their discipline called macroeconomics. Yet theory in this field has not developed through creation of appropriate micro-to-macro transformations, but via a short cut, using the idea of a representative agent. If I am correct, the deficiencies of macroeconomics as a policy science lie in this substitution.

[1] I have not attempted to list systematically the ways that social structure can enter into economic models. One way, exemplified in assortative mating markets, is through other actors (or their attributes) entering into an actor's utility function. A second, illustrated by trust, is in expectations about other actors' behavior. A third lies in differential communication patterns. But there are very likely others as well.

REFERENCES

Ashton, T. S., "The Bill of Exchange and Private Banks in Lancashire, 1790–1830," *Economic History Review*, No. 1–2, 1945, *15*, 25–35.

Clark, Rodney, *The Japanese Company*, New Haven: Yale University Press, 1979.

Embree, John F., *Suze Mura: A Japanese Village*, Chicago: University of Chicago Press, 1939.

Mackay, Charles, *Memoirs of Extraordinary Popular Delusions and the Madness of Crowds*, London: 1892.

Schelling, Thomas, *Micromotives and Macrobehavior*, New York: Norton, 1978.

Against Parsimony: Three Easy Ways of Complicating Some Categories of Economic Discourse

By ALBERT O. HIRSCHMAN*

In his well-known article on "Rational Fools," Amartya Sen asserted that "traditional [economic] theory has *too little* structure" (1977, p. 335). Like any virtue, so he seemed to say, parsimony in theory construction can be overdone and something is sometimes to be gained by making things *more complicated*. I have increasingly come to feel this way. Some years ago, I suggested that criticism from customers or "voice" should be recognized as a force keeping management of firms and organizations "on their toes," alongside with competition or "exit," and it took a book (1970) to cope with the resulting complications. Here I deal with various other realms of economic inquiry that stand similarly in need of being rendered more complex. In concluding I examine whether the various complications have some element in common—that would in turn simplify and unify matters.

I. Two Kinds of Preference Changes

A fruitful distinction has been made, by Sen and others, between first- and second-order preferences, or between preferences and metapreferences, respectively. I shall use the latter terminology here. Economics has traditionally dealt only with (first-order) preferences, that is, those that are *revealed* by agents as they buy goods and services. But the concept of metapreference must be of concern to the economist, to the extent that he claims an interest in understanding processes of economic *change*. Its starting point is a very general observation on *human nature*: men and women have the ability to step back from their "revealed" wants, volitions, and preferences, to ask themselves whether they really want these wants and prefer these preferences, and consequently to form metapreferences that may differ from their preferences. Unsurprisingly, it is a philosopher, Harry Frankfurt (1971), who first put matters this way. He argued that this ability to step back is unique in humans, but is not present in all of them. Those who lack this ability he called "wantons": they are entirely, unreflectively in the grip of their whims and passions.

As I have pointed out before (1982, p. 71), certainty about the existence of metapreferences can only be gained through *changes* in preferences, that is, through changes in actual choice behavior. If preferences and metapreferences always coincide so that the agent is permanently at peace with himself no matter what choices he makes, then the metapreferences hardly lead an independent existence and are mere shadows of the preferences. If, on the other hand, the two kinds of preferences are permanently at odds so that the agent always acts against "his better judgment," then again the metapreference can not only be dismissed as wholly ineffective, but doubts will arise whether it is really there at all.

Changes in choice behavior are therefore essential for validating the concept of metapreferences; conversely, this concept is useful in illuminating the varied nature of preference change, for it is now possible to distinguish between two kinds of *preference changes*. One is the reflective kind, preceded as it is by the formation of a metapreference that is at odds with the observed and hitherto practiced preference. But there are also preference changes that take place without any elaborate antecedent development of metapreferences. Following Frankfurt's terminology, the unreflective changes in preferences might be called "wanton." These are the

*Professor of Social Science, The Institute for Advance Study, Princeton, NJ 08540. A more complete version of this paper will appear in *Bulletin of the American Academy of Arts and Sciences*, forthcoming.

preference changes economists have primarily focused on: haphazard, publicity-induced, and generally minor (apples vs. pears) *change in tastes*. In contrast, the nonwanton change of preference is not really a change in tastes at all. A taste is almost *defined* as a preference about which you do not argue—de gustibus non est disputandum. A taste about which you argue, with others *or yourself*, ceases ipso facto being a taste—it turns into a *value*. When a change in preferences has been preceded by the formation of a metapreference much argument has obviously gone on within the divided self; it typically represents a *change in values* rather than a change in tastes.

Given the economists' concentration on, and consequent bias for, wanton preference changes, changes of the reflective kind have tended to be downgraded to the wanton kind by assimilating them to changes in *tastes*: thus patterns of discriminatory hiring have been ascribed to a "taste for discrimination" (Gary Becker, 1957) and increases in protectionism have similarly been analyzed as reflecting an enhanced "taste for nationalism" (Harry Johnson, 1965). Such interpretations strike me as objectionable on two counts: first, they impede a serious intellectual effort to understand what are strongly held values and difficult-to-achieve changes in values rather than tastes and changes in tastes; second, the illusion is fostered that "raising the cost" of discrimination (or nationalism) is the simple and sovereign policy instrument for getting people to indulge less in those odd "tastes."

In the light of the distinction between wanton and nonwanton preference changes, or between changes in tastes and changes in values, it also becomes possible to understand—and to criticize—the recent attempt of Becker and George Stigler (1977) to do without the notion of preference changes for the purpose of explaining changes in behavior. Equating preference changes to changes in what they themselves call "inscrutable, often capricious tastes" (p. 76), they find, quite rightly, any changes in those kinds of tastes (our wanton changes) of little analytical interest. But in their subsequent determination to explain all behavior change through price and income differences, they neglect one important source of such change: autonomous, reflective change in values. For example, in their analysis of beneficial and harmful addiction they take the elasticity of the individual's demand curve for music or heroin as given and, it would seem, immutable. May I urge that changes in values do occur from time to time in the lives of individuals, of generations, and from one generation to another, and that those changes and their effects on behavior are worth exploring—that, in brief, de valoribus *est* disputandum?

II. Two Kinds of Activities

From consumption I now turn to production and to human activities such as work and effort involved in achieving production goals. Much of economic activity is directed to the production of (private) goods and services that are then sold in the market. From the point of view of the firm, the activity carries with it a neat distinction between process and outcome, inputs and outputs, or costs and revenue. From the point of view of the individual participant in the process, a seemingly similar distinction can be drawn between work and pay or between effort and reward. Yet there is a well-known difference between the firm and the individual: for the firm any outlay is unambiguously to be entered on the negative side of the accounts whereas work can be more or less irksome or pleasant—even the same work can be felt as more pleasant by the same person from one day to the next. This problem, in particular its positive and normative consequences for income differentials, has attracted the attention of a long line of economists starting with Adam Smith. Most recently Gordon Winston has distinguished between "process utility" and "goal utility" (1982, pp. 193–97). While such a distinction makes it clear that the means to the end of productive effort need not be entered on the negative side in a calculus on satisfaction, it keeps intact the basic instrumental conception of work, the means-end dichotomy on which our understanding of the work and production process has been essentially—and, up to a point, so usefully—based.

But there is need to go further if the complexity and full range of human activities, productive and otherwise, are to be appreciated. Once again, more structure would be helpful. The possible existence of wholly *noninstrumental* activities is suggested by everyday language: it speaks of activities that are undertaken "for their own sake" and that "carry their own reward." These are somewhat trite, unconvincing phrases: after all, any sustained activity, with the possible exception of pure play, is undertaken with some idea about an intended outcome. A person who claims to be working exclusively for the sake of the rewards yielded by the exertion itself is usually suspect of hypocrisy: one feels he is "really" after the money, the advancement or—at least—the glory, and thus is an instrumentalist after all.

Some progress can be made with the matter by looking at the varying predictability of the intended outcome of different productive activities. Certain activities, typically of a routine character, have perfectly predictable outcomes. With regard to such tasks, there is no doubt in the individual's mind that effort will yield the anticipated outcome—an hour of labor will yield the well-known, fully visualized result as well as entitle the worker, if he has been contracted for the job, to a wage that can be used for the purchase of desired (and usually also well-known) goods. Under these conditions, the separation of the process into means and ends, or into costs and benefits, occurs almost spontaneously and work assumes its normal instrumental character.

But there are many kinds of activities, from that of a research and development scientist to that of a composer or an advocate of some public policy, whose intended outcome cannot be relied upon to materialize with certainty. Among these activities there are some—applied laboratory research may be an example—whose outcome cannot be predicted for any single day or month; nevertheless, success in achieving the intended result steadily gains in likelihood as the period during which work is carried on gets longer. In this case, the uncertainty is of a probabilistic nature and one can speak of a certainty equivalent with regard to the output of the activity in any given period so that, once again, the separation of the process into means and ends is being experienced and work of this sort largely retains its instrumental cast.

I now come to a more puzzling kind of nonroutine activities. From their earliest origins, men and women appear to have allocated a considerable portion of their time to undertakings whose success is simply unpredictable. These are activities such as the pursuit of truth, beauty, justice, liberty, community, friendship, love, salvation, and so on. As a rule, these pursuits are of course carried on through a variety of exertions for apparently limited and specific objectives (writing a book, participating in a political campaign, etc.). Nevertheless, an important component of the activities thus undertaken is best described not as labor or work, but as *striving*—a term that precisely intimates the lack of a reliable relation between effort and result. A means-end or cost-benefit calculus is impossible under the circumstances.

The question now arises why such activities should be taken up at all, as long as their successful outcome is so wholly uncertain. Moreover, they certainly are not always pleasant in themselves—in fact some of them can be quite strenuous or highly dangerous. Do we have here then another paradox or puzzle, one that relates not just to voting (why do "rational" people bother to vote?), but to a much wider and most vital group of human activities? I suppose we do—from the point of view of instrumental reason, noninstrumental action is bound to be something of a mystery. But I have proposed an at least half-rational explanation: these noninstrumental activities whose outcome is so uncertain are strangely characterized by a certain fusion of (and confusion between) striving and attaining (see my 1982 book, pp. 84–91). He who strives after truth (or beauty) frequently experiences the conviction, fleeting though it may be, that he has found (or achieved) it. He who participates in a movement for liberty or justice frequently has the experience of already bringing these ideals within reach. In Pascal's formulation: "The hope Christians have to possess an infinite good is mixed with actual enjoyment...for they are not like people who would hope for a kingdom of which they, as subjects, have nothing; rather, they hope for holiness, and for freedom from injustice, and they partake of both" (*Pensées*, 540, Brunschvicg edition, my translation).

This fusion of striving and attaining is a fact of experience that goes far in accounting for the existence and importance of noninstrumental activities. As though in compensation for the uncertainty about the outcome, the striving effort is colored by the goal and in this fashion makes for an experience that is very different from merely agreeable, pleasurable or even "stimulating": in spite of its frequently painful character it has a well-known "intoxicating" quality.

The foregoing interpretation of noninstrumental action is complemented by an alternative view which has been proposed by the sociologist Alessandro Pizzorno (1983). For him, participation in politics is often engaged in because it enhances one's feeling of belonging to a group. I would add that noninstrumental action in general makes you feel more like a "real person." Such action can then be considered, in economic terms, as an *investment in individual and group identity.* The feeling of having achieved belongingness and personhood may of course be just as evanescent as the fusion of striving and attaining to which I referred earlier. The two views are related attempts at achieving an uncommonly difficult insight: to think instrumentally about the noninstrumental.

But why should economics be concerned with all this? Is it not enough for this discipline to attempt an adequate account of man's instrumental activities—a vast area indeed—while leaving the other, somewhat murky regions alone? Up to a point such a limitation made sense. But as economics has grown more ambitious, it becomes of increasing importance to appreciate that the means-end, cost-benefit model is far from covering all aspects of human activity and experience. Take the analysis of political action, an area in which economists have become interested as a natural extension of their work on public goods. Here the neglect of the noninstrumental mode of action was responsible for the inability of the "economic approach" to understand why people bother to vote and why they engage from time to time in collective action.

Once the noninstrumental mode is being paid some attention it becomes possible to account for these otherwise puzzling phenomena. It is the fusion of striving and attaining, characteristic of noninstrumental action, that led me to a conclusion exactly opposite to the "free ride" argument with respect to collective action:

> since the output and objective of collective action are...a public good available to all, the only way an individual can raise the benefit accruing to him from the collective action is by stepping up *his own input*, his effort on behalf of the public policy he espouses. Far from shirking and attempting to get a free ride, a truly maximizing individual will attempt to be as activist as he can manage,.... [1982, p. 86]

The preceding argument does not imply, of course, that citizens will never adopt the instrumental mode of action with respect to action in the public interest. On the contrary, quite a few of them may well move from one mode to the other, and such oscillations could help explain the observed instability both of individual commitment and of many social movements in general.

A better understanding of collective action is by no means the only benefit that stands to flow from a more open attitude toward the possibility of noninstrumental action. As has been argued earlier, a strong affinity exists between instrumental and routine activities, on the one hand, and between noninstrumental and nonroutine activities, on the other. But just as I noted the existence of nonroutine activities that are predominantly instrumental (in the case of an applied research laboratory), so can routine work have more or less of a noninstrumental component, as Veblen stressed in *The Instinct of Workmanship*. Lately the conviction has gained ground that fluctuations in this component must be drawn upon to account for variations in labor productivity and for shifts in industrial leadership. It does make a great deal of difference, so it seems, whether people look at their work as "just a job" or also as part of some collective celebration.

III. "Love": Neither Scarce Resource Nor Augmentable Skill

My next plea for complicating economic discourse also deals with the production side, but more specifically with the role of one important prerequisite or ingredient known variously as morality, civic spirit, trust, observance of elementary ethical norms, and so on. The need of any functioning economic system for this "input" is widely recognized. But disagreement exists over what happens to this input as it is being used.

There are essentially two opposite models of factor use. The traditional one is constructed on the basis of given, depletable resources that get incorporated into the product. The scarcer the resource the higher its price and the less of it will be used by the economizing firm in combination with other inputs. A more recent model recognizes the possibility of "learning by doing" (Kenneth Arrow, 1962). Use of a resource such as a skill has the immediate effect of improving the skill, of enlarging (rather than depleting) its availability. The recognition of this sort of process was a considerable, strangely belated insight.

Because the "scarce resource" model has long been dominant, it has been extended to domains where its validity is highly dubious. Some thirty years ago, Dennis Robertson wrote a characteristically witty paper entitled "What Does the Economist Economize?" (1956). His often cited answer was: love, which he called "that scarce resource" (p. 154). Robertson explained, through a number of well-chosen illustrations from the contemporary economic scene, that it was the economist's job to create an institutional environment and pattern of motivation where as small a burden as possible would be placed, for the purposes of society's functioning, on this thing "love," a term he used as a shortcut for morality and civic spirit. In so arguing, he was of course at one with Adam Smith who celebrated society's ability to do without "benevolence" (of the butcher, brewer, and baker) as long as individual "interest" was given full scope. Robertson does not invoke Smith, quoting instead a telling phrase by Marshall: "Progress chiefly depends on the extent to which the *strongest* ınd not merely the *highest* forces of human nature can be utilized for the increase of social good" (p. 148). This is yet another way of asserting that the social order is more secure when it is built on interest rather than on love or benevolence. But the sharpness of Robertson's own formulation makes it possible to identify the flaw in this recurrent mode of reasoning.

Once love and particularly public morality is equated to a scarce resource, the need to economize it seems self-evident. Yet a moment's reflection is enough to realize that the analogy is not only questionable, but a bit absurd—and therefore funny. Take, for example, the well-known case of the person who drives in the morning rush hour and quips, upon yielding to another motorist: "I have done my good deed for the day; for the remainder, I can now act like a bastard." What strikes one as funny and absurd here is precisely the assumption, on the part of our driver, that he comes equipped with a strictly limited supply of good deeds; that, in other words, love should be treated as a scarce resource—just as Robertson claimed. We know instinctively that the supply of such resources as love or public spirit is not fixed or limited as may be the case for other factors of production. The analogy is faulty for two reasons: first of all, these are resources whose supply may well increase rather than decrease through use; second, these resources do not remain intact if they stay unused; like the ability to speak a foreign language or to play the piano, these moral resources are likely to become depleted and to atrophy if *not* used.

In a first approximation, then, Robertson's prescription appears to be founded on a confusion between the *use of a resource* and *the practice of an ability*. While human abilities and skills are valuable economic resources, most of them respond positively to practice, in a learning-by-doing manner, and negatively to nonpractice. It was on the basis of this atrophy dynamic that the U.S. system for obtaining an adequate supply of human blood for medical purposes, with its only partial reliance on voluntary giving, has been criticized by Richard Titmuss, the British sociologist. And a British political economist, Fred Hirsch (1976), has generalized the point: once a social system, such as capitalism, convinces everyone that it can dispense with morality and public spirit, the universal pursuit of self-interest being all that is needed for satisfactory performance, the system will undermine its own viability which is in fact premised on civic behavior and on the respect of certain moral norms to a far greater extent than capitalism's official ideology avows.

How is it possible to reconcile the concerns of Titmuss-Hirsch with those seemingly opposite, yet surely not without some foundation, of Robertson, Adam Smith, and Alfred Marshall? The truth is that, in his fondness for paradox, Robertson did his position a disservice: he opened his flank to easy attack when he equated love to some factor of production in strictly limited supply that needs to be economized. But what about the alternative analogy that equates love, benevolence, and public spirit to a skill that is improved through practice and atrophies without it? This one, too, has its weak points. Whereas public spirit will atrophy if too few demands are made upon it, it is not at all certain that the practice of benevolence will indefinitely have a positive feedback effect on the supply of this "skill." The practice of benevolence yields satisfaction (makes you feel good), to be sure, and therefore feeds upon itself up to a point, but this process is very different from practicing a manual (or intellectual) skill: here the practice leads to greater *dexterity* which is usually a net addition to one's abilities, that is, it is not acquired at the expense of some other skill or ability. In the case of benevolence, on the other hand, the point is soon reached where increased practice does conflict with self-interest and even self-preservation: our quipping motorist, to go back to him, has not exhausted his daily supply of benevolence by yielding once, but there surely will be *some* limit to his benevolent driving behavior, in deference to his own vital—perhaps ethically compelling—displacement needs.

Robertson had a point, therefore, when he maintained that there could be institutional arrangements which make *excessive* demands on civic behavior just as Titmuss and Hirsch were right in pointing to the opposite danger: the possibility, that is, that society makes *insufficient* demands on civic spirit. In both cases, there is a shortfall in public spirit, but in the cases pointed to by Robertson, the remedy consists in institutional arrangements placing *less* reliance on civic spirit and more on self-interest whereas in the situations that have caught the attention of Titmuss and Hirsch there is need for *increased* emphasis on, and practice of, community values and benevolence. These two parties argue along exactly opposite lines, but both have a point. Love, benevolence, and civic spirit are neither scarce factors in fixed supply, nor do they act like skills and abilities that improve and expand more or less indefinitely with practice. Rather, they exhibit a complex, composite behavior: they atrophy when not adequately practiced and appealed to by the ruling socioeconomic regime, yet will once again make themselves scarce when preached and relied on to excess.

To make matters worse, the precise location of these two danger zones—which, incidentally, correspond roughly to the complementary ills of today's capitalist and centrally planned societies—is by no means known, nor are these zones ever stable. An ideological-institutional regime that in wartime or during some other time of stress and public fervor is ideally suited to call forth the energies and efforts of the citizenry is well advised to give way to another that appeals more to private interest and less to civic spirit in a subsequent, less exalted period. Inversely, a regime of the latter sort may, because of the ensuing "atrophy of public meanings" (Charles Taylor, 1970, p. 123), give rise to anomie and unwillingness ever to sacrifice private or group interest to the public weal so that a move back to a more community-oriented regime would be called for.

IV. Conclusion

I promised, earlier on, to inquire whether the various complications of traditional concepts that have been proposed have any common structure. The answer should be obvious: all these complications flow from a single source—the incredible complexity of human nature which was disregarded by traditional theory for very good reasons, but which must be spoon-fed back into the traditional findings for the sake of greater realism.

A plea to recognize this complexity was implicit in my earlier insistence that "voice" be granted a role in certain economic processes alongside "exit," or competition. The efficient economic agent of traditional theory is essentially a silent scanner and "superior statistician" (Arrow, 1978) whereas I argued that she also has considerable gifts of verbal and nonverbal communication and persuasion that will enable her to affect economic processes.

Another fundamental characteristic of humans is that they are *self-evaluating* beings, perhaps the only ones among living organisms. This simple fact forced the intrusion of metapreferences into the theory of consumer choice and made it possible to draw a distinction between two fundamentally different kinds of preference changes. The self-evaluating function could be considered a variant of the communication or voice function: it also consists in a person addressing, criticizing, or persuading someone, but this someone is now the *self* rather than a supplier or an organization to which one belongs. But let us beware of excessive parsimony!

In addition to being endowed with such capabilities as communication, persuasion and self-evaluation, man is beset by a number of fundamental, unresolved, and perhaps unresolvable tensions. A tension of this kind is that between instrumental and noninstrumental modes of behavior and action. Economics has, for very good reasons, concentrated wholly on the instrumental mode. I plead here for a concern with the opposite mode, on the grounds 1) that it is not wholly impervious to economic reasoning; and 2) that it helps us understand matters that have been found puzzling, such as collective action and shifts in labor productivity.

Finally I have turned to another basic tension man must live with, this one resulting from the fact that he lives in society. It is the tension between self and others, between self-interest, on the one hand, and public morality, service to community, or even self-sacrifice, on the other, or between "interest" and "benevolence" as Adam Smith put it. Here again economics has concentrated overwhelmingly on one term of the dichotomy, while putting forward simplistic and contradictory propositions on how to deal with the other. The contradiction can be resolved by closer attention to the special nature of public morality as an "input."

In sum, I have complicated economic discourse by attempting to incorporate into it two basic human endowments and two basic tensions that are part of the human condition. To my mind, this is just a beginning.

REFERENCES

Arrow, Kenneth J., "The Economic Implications of Learning by Doing," *Review of Economic Studies*, June 1962, *29*, 155–73.

______, "The Future and the Present in Economic Life," *Economic Inquiry*, April 1978, *16*, 160.

Becker, Gary S., *The Economies of Discrimination*, Chicago: Chicago University Press, 1957.

______ **and Stigler, George,** "De Gustibus Non Est Disputandum," *American Economic Review*, March 1977, *67*, 76–90.

Frankfurt, Harry G., "Freedom of the Will and the Concept of a Person," *Journal of Philosophy*, January 1971, *68*, 5–20.

Hirsch, Fred, *Social Limits to Growth*, Cambridge: Harvard University Press, 1976.

Hirschman, Albert O., *Exit, Voice, and Loyalty*, Cambridge: Harvard University Press, 1970.

______, *Shifting Involvements: Private Inter-*

est and Public Action, Princeton: Princeton University Press, 1982.

Johnson, Harry G., "A Theoretical Model of Economic Nationalism in New and Developing States," *Political Science Quarterly*, June 1965, *80*, 169–85.

Pizzorno, Alessandro, "Sulla razionalità della scelta democratica," *Stato e Mercato*, April 1983, No. 7, 3–46.

Robertson, Dennis H., "What Does the Economist Economize?," in *Economic Commentaries*, London: Staples Press, 1956, 147–55.

Sen, Amartya K., "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory," *Philosophy and Public Affairs*, Summer 1977, *6*, 317–44.

Taylor, Charles, *The Pattern of Politics*, Toronto: McClelland and Stewart, 1970.

Winston, Gordon C., *The Timing of Economic Activities*, Cambridge: Cambridge University Press, 1982.

KEYNES, SCHUMPETER, MARX CENTENARY SESSION

Keynes and Economics Today

By DON PATINKIN*

At the end of a centenary year in which there have been endless speculations about what John Maynard Keynes might have said about economic thought and policy today, I am not sure that there is much new to be said on this question. Indeed, I have some doubts as to its meaningfulness. Nevertheless, I shall attempt to offer some general observations about it.[1]

To speculate about this question is, first of all, to ask how the discipline of economics today (and macroeconomics in particular) differs from that of the 1930's. With respect to the state of economics in general, there are four main characteristics that I would list: the much greater use of mathematics and the related greater rigor of the analysis; the much greater emphasis on empirical work, and on the econometric testing of hypotheses in particular; the emphasis given today to the economics of growth in developed, and especially underdeveloped, countries; and, more generally, the much greater degree of specialization in economics and the extension of its analysis to many additional aspects of life.

There is little, if anything, in Keynes' writings that provides a basis for speculation about these last two characteristics. Like practically all economists of the interwar depression period, Keynes was not concerned with either growth (indeed, in 1937 he delivered a lecture on "Some Consequences of a Declining Population," *JMK*, XIV, pp. 124–33) or the underdeveloped countries. True, Keynes' first book—published in 1913—was devoted to *Indian Currency and Finance*. But its concern was not with India as such (which he never visited), but with what the Western world could learn from India's experience with the sterling exchange standard. In a similar way, I think Keynes would have concerned himself today with the impact on the financial system of the Western world of the tremendous liquid holdings of the OPEC countries, as well as of the recent debt crises of some of the developing countries.

With respect to the increased use of mathematical analysis, we have Keynes' oft-cited criticism in the *General Theory* of "symbolic pseudomathematical methods of formalizing a system of economic analysis...which allow the author to lose sight of the complexities and interdependencies of the real world in a maze of pretentious and unhelpful symbols" (*GT*, pp. 297–98). Let us, however, not take this statement too seriously. First of all, Keynes' own analysis in his earlier *Treatise on Money* (1930) was, in fact, largely based on fairly mechanical applications of the so-called fundamental equations. Indeed, if ever an author made use of "a maze of pretentious and unhelpful symbols" that author was Keynes of the *Treatise*

*The Hebrew University of Jerusalem and the Maurice Falk Institute for Economic Research, Israel. Without burdening them with any responsibility, I thank Stanley Fischer, Peter Howitt, David Laidler, and Rudy Dornbusch for helpful comments on an earlier draft. An extended version of this paper will appear early 1985 in a volume edited by Claus Köehler and Jochem Langkau, *Monetary and Fiscal Policy in the Eighties*, to be published by Verlag Neue Gesellschaft, Bonn.

[1] All references to the writings of Keynes are to the new edition of his *Collected Writings*. For simplicity, I shall refer to the two volumes of his *Treatise on Money* by the short titles *Treatise*, I (II) or *TM*, I (II), respectively. The *General Theory* will sometimes be further abbreviated to *GT*. Specific volumes in Keynes' *Collected Writings* will be referred to, for example, as *JMK*, IX; *JMK*, XIII, and so forth. I have in the following drawn freely on the material in my *Keynes' Monetary Thought* (1976) and in chapter 9 of my *Anticipations of the General Theory? And Other Essays on Keynes* (1982). These books will be referred to henceforth as *KMT* and *Anticipations*, respectively.

(see *KMT*, chs. 4–7). Furthermore, I strongly suspect that a comparison of the *General Theory* (and a fortiori the *Treatise*) with the other works on economic theory that were written during that period would actually show Keynes' works to be among the more mathematical of them. Indeed, in his review of the *General Theory*, Austin Robinson commented that "even for the ordinary economist, the argument, being largely in mathematical form, is difficult" (1936, p. 472). At the same time, I am sure that Keynes would have had little patience with what is now termed "highbrow mathematical economics," and much of the "middlebrow" as well. In 1944, Keynes agreed (after politely protesting that he had not done much in the way of econometrics) to accept the presidency of the Econometric Society; looking at the contents of *Econometrica* over the past decade and more, I think that today he would have persisted with his protests.

What would Keynes have thought of present-day econometrics? Here there immediately comes to mind his famous 1939 review of Tinbergen's work, and let me emphasize that this review was devoted not to the much better known second volume of this study on business cycles in the United States, but to the first volume on the principles of multiple correlation analysis. Accordingly, the criticisms Keynes presented in this review were leveled not at Tinbergen's pioneering and (certainly for that time) ambitious forty-six equation model of the U.S. economy, but at the use of regression analysis to estimate even a single equation! Nor can we dismiss these criticisms, for some of them dealt with methodological issues which were subsequently to become basic concerns of the econometric literature. Thus, to use current terminology (which Keynes obviously did not), he criticized Tinbergen's work for being guilty of both the specification bias and the simultaneous-equation bias. Similarly, Keynes emphasized the basic difficulty of measuring expectations (*Anticipations*, pp. 227–30).

Keynes concluded his review of Tinbergen's work by saying: "I have a feeling that Professor Tinbergen may agree with much of my comment, but his reaction will be to engage another ten computers and drown his sorrows in arithmetic" (*JMK*, XIV, p. 318). And if this is what Keynes said at a time when one had to think twice before undertaking the laborious and time-consuming task of estimating a single multiple-regression equation on a mechanical desk calculator, what would he say today, in this age of instant estimation?

In contrast with his attitude toward econometric models, Keynes would undoubtedly have been an enthusiastic supporter of a second and related postwar development in empirical economics: namely the great development of macroeconomic statistics. Throughout the interwar period, Keynes continued to complain about the inadequacy of such statistics. And though in the early 1930's he did not give Colin Clark's work on national-income estimates the support he should have, Keynes played a crucial role in the subsequent development of national-income accounting in Britain in World War II (*Anticipations*, pp. 248–54).

Let me turn now to Keynes' possible views on postwar developments in macroeconomics. I begin with monetarism. As we all know, the policy conclusion of the *General Theory* was that the problem of unemployment could not be solved by monetary policy alone, but that a program of government investment (expenditures on public works and the like) was necessary. Let me first emphasize that what Keynes meant by monetary policy is not what today's monetarists mean by it. For what Keynes always meant by this term was not a policy of controlling the quantity of money, but one of controlling the rate of interest: lowering it by means of central bank open market purchases to ward off deflation and unemployment, and raising it to ward off inflation. Clearly such central bank operations generate changes in the quantity of money; but this quantity was not Keynes' target variable.

What would have been Keynes' view of monetarism? If by this term we mean the view that fiscal policy has no role to play in a contracyclical policy, then I have no doubt that he would have rejected it. The legend that Keynes in the last years of his life had second thoughts about the desirability of fiscal policy and in this context said "I am not a Keynesian" is such an irresistible one—the classic motif of the sinner repentant on his death bed—that I hate to spoil it by noting that it has never really been corroborated and is indeed contradicted by Keynes' writings in the years after the *General Theory* (*Anticipations*, p. 214).

To return to the question of monetarist policy, I feel that Keynes would have rejected it on theoretical and empirical grounds similar to those on which he rejected interest rate policy in the *General Theory* (especially ch. 19) and in other writings: namely that the causal mechanism through which monetarist policy was supposed to work had not been convincingly specified; and that in any event experience has shown that this policy could not be depended upon to assure full employment.

At the same time, let me say that to my mind the economist closest to Keynes in his conception of the role of the economist in society is none other than the leading spokesman of monetarism, Milton Friedman. For both of these men the purpose of economic analysis is not only to construct theoretical models, but to lead to policy recommendations—and accordingly both had a continuous and detailed concern with the current empirical data on the workings of the economy. Furthermore, both regarded as an essential part of their task as economists not only to formulate policy positions, but to generate public opinion in support of them. And both sedulously exploited all means of communication for this purpose: articles in leading newspapers and magazines; books and pamphlets; participation in radio and television programs (of course, there was no television in Keynes' day; but can anyone doubt that he would have been a television personality if there had been?); appearances and testimony before government committees and bodies; and personal contacts with leading government officials responsible for formulating and carrying out policy, while at the same time eschewing (except in wartime) official government positions. And both tried to achieve their respective goals by tirelessly hammering away at a single, oversimplified theme: public expenditures in the case of Keynes; money supply in the case of Friedman. Yet another common characteristic is that for both of them the data always provided the right answer.

Before leaving this subject, I would like to emphasize that Keynes also shared with today's monetarists a great concern with avoiding inflation. It is not surprising that this was Keynes' view during the period of the disastrous hyperinflations that followed World War I; but even in his 1930 *Treatise*, after more than five years of deflation and unemployment in Britain, Keynes continued to be concerned with the danger of inflation. Perhaps because of the even deeper depression that Britain had fallen into at the time of the *General Theory*, this book is little concerned with the problem of inflation. It is, however, highly significant that after Britain began its rearmament program in early 1937—and when unemployment was still around 12 percent—Keynes expressed concern with the possible inflationary outcome of such a program that might be generated by the immobility of labor (Terence Hutchison, 1977, pp. 10–14). And once war broke out, Keynes wrote his *How to Pay for the War* (1940), whose major purpose was to present a specific program for financing the war without generating inflation.

I must also emphasize that a recurrent theme of the *General Theory* (pp. 173, 249, 253, 296, and 301) is that as the level of employment in an economy increased as a result of an increase in effective demand, the money wage rate would begin to rise even before full employment was reached. In this theme I see an adumbration of one aspect of the Phillips-curve analysis, namely the coexistence of inflation and unemployment. But my main point here is that I infer from Keynes' view on this matter that he would not have considered such a coexistence to be a contradiction to his theory. And I also conjecture that it would not have surprised him to see the phenomenon of inflation *cum* unemployment manifest itself even more strongly in the postwar world, with its increased economic and political power of labor and its lessened faith in the antiinflationary proclamations of governments.

How would Keynes have suggested dealing with this phenomenon? Here, as in some other contexts, he would probably have advocated making some institutional arrangement to deal with the problem: in this case, an institutional arrangement requiring the cooperation of labor unions, industry, and government to determine the general level of money wages. But I suspect that he would have been quite skeptical about the possibility of creating and effectively carrying out such an institutional arrangement—just as in his 1930 *Treatise* (II, pp. 336–37), he was skeptical about the possibility of achieving the cooperation of the central banks of Britain, the United States, and France in order to deal with the problem of unemployment that then existed.

Let me turn now to Keynes' possible views about the "new macroeconomics." Of one thing we can be sure: he would not have been happy to learn that the title of one of the early and influential contributions to this literature was entitled "After Keynesian Macro-economics." Nor would he have drawn much comfort from the fact that a defense of his theory was entitled "How Dead is Keynes?" But let me turn to the substantive issues involved, distinguishing between two main components of the new macroeconomics: the assumption of rational expectations and the assumption of market equilibrium.

With reference to the first of these, I must again distinguish between rational expectations in the short run and in the long run. I think that Keynes would have been willing to accept this assumption within a short-run context. Thus, in notes which he prepared for the lectures on monetary theory that he gave at Cambridge in 1937, he wrote "... the theory of effective demand is substantially the same if we assume that short-period expectations are always fulfilled ..." (*JMK*, XIV, pp. 180–81). Frankly, I am not sure how this statement should be interpreted, but it does seem to indicate a willingness to accept the assumption of rational expectations in a short-run context.

The situation is entirely different with respect to the long-run expectations involved in, for example, investment decisions. Thus in chapter 12 of the *General Theory* on "The State of Long-Term Expectations," Keynes emphasized that, in view of the great uncertainty about the future, "our decisions to do something positive ... can only be taken as a result of animal spirits ... and not as an outcome of a weighted average of quantitative benefits multiplied by quantitative probabilities" (*GT*, p. 161). And, at an earlier point in this chapter (*GT*, p. 148, fn. 1), he explained that "by 'very uncertain' I do not mean the same thing as 'very unprobable'," and refers to chapter 6 of his *Treatise on Probability* (1921). Similarly, in his 1937 *Quarterly Journal of Economics* article, Keynes emphasizes that the uncertainty which characterizes so much of economic life is one for which "there is no scientific basis on which to form any calculable probability whatever" (*JMK*, XIV, p. 114). Clearly Keynes' view about the absence of a probability calculus in this context is inconsistent with the basic point of departure of the rational expectations approach.

Insofar as the assumption of market equilibrium is concerned, here there can be no doubt that Keynes would have rejected it. Indeed, the central policy message of the *General Theory* is that

> There is, therefore, no grounds for the belief that a flexible wage policy is capable of maintaining a state of continuous full employment; any more than for the belief that an open-market monetary policy is capable, unaided, of achieving this result. The economic system cannot be made self-adjusting along these lines. [*GT*, p. 267]

And this lack of faith in the efficacy of the market-equilibrium process in a macroeconomic context also manifests itself in such earlier writings as *The Economic Consequences of Mr. Churchill* (1925; *JMK*, IX, pp. 227–29, et passim) and the *Treatise* (I, pp. 141, 151, 244–45, 265). Nor would Keynes have been impressed by the contention sometimes advanced that the market would not permit a situation of unemployment to persist because contracts could then be made which would make everyone better off. Indeed, I would conjecture that, as one who had seen how the most civilized countries of the world had engaged for four long years of stalemated trench warfare in the mutual slaughter of the best of their young men, Keynes was not predisposed to believe in natural forces that always brought agents to generate a mutually beneficial situation. Because of the uncertainty of how others react to our actions, the actual world for Keynes was one that—in a macro context—could readily lead to the "globally irrational" results of the prisoner's dilemma; not to the rational results of the Walrasian auctioneer.

I cannot conclude this paper without some conjectures about what Keynes would have said about the many interpretations of the *General Theory* that have been presented since his death. What would he have said of an interpretation of the *General Theory* which claims that it was based on the assumption of rigid money wages, when he had deliberately structured the argument of the book to lead up to chapter 19 on "Changes in Money-Wages" (which he prefaced with the observation that "it was not possible ... to discuss this matter fully until our own theory had been developed" *GT*, p. 257), in which he applied his theory to show why the "classical" remedy of a reduction in the level of money wages could not be depended upon to increase the level of employment? What would he have said of an interpretation which claims that the argument of the *General Theory* is based on the "liquidity trap," when he had explicitly stated that "whilst this limiting case might become practically important in the future, I know of no example of it hitherto" (*GT*, p. 207)? What would he have said of an interpretation of the *General Theory* which sees as its main achievement the analysis of the determination of the price level, particularly in times of inflation—despite the fact that Keynes wrote his book in a period of depression and deflation, and emphasized that "the theory of prices falls into its proper place as a matter which is subsidiary to our general theory" (*GT*, p. 32)? What would he have said of an interpretation of the *General Theory* which sees as its central message the fact that economic decisions (and particularly those relating to investment) are made in a world of uncertainty —when despite the undeniably very important role of uncertainty in the book, Keynes did not refer to it explicitly either in his introductory chapter 3 on "The Principles of Effective Demand," in which he presented "a brief summary of the theory of employment to be worked out in the course of the following chapters" (*GT*, p. 27), or in his summary chapter 18 on "The General Theory of Employment Re-Stated"? What would he have said of an interpretation which, in order to recreate the *General Theory* in the image of Cambridge (England) of today, claims that this book rejected marginal analysis, when—in contrast to the *Treatise*, in which the word "marginal" does not appear—Keynes of the *General Theory* emphasized in its Preface that the analysis of the book was "linked up with our fundamental theory of value" and accordingly applied marginal analysis in his treatment of the markets for labor, consumer goods, and investment goods, respectively? Finally, what would he have said of an attempt to distinguish between "Keynes and the Keynesians" (here is another irresistible, classic theme: the good Czar, whose attempts to improve the lot of his people are, unknown to him, being thwarted by his evil ministers) —despite the fact that he had "next to nothing to say by way of criticism" of Hicks' *IS-LM* interpretation, which was to become the standard one of "the Keynesians" (*JMK*, XIV, p. 79)?[2]

What would he have said about all these exegetical attempts? I suspect that he would have repeated—while raising it to the *n*th power—what he said in concluding a long and tiresome correspondence in 1938 on a note that someone had sent him on an aspect of the *General Theory*: "... the enclosed, as it stands, looks to me more like theology than economics! ... I am really driving at some-

[2] The fact that in recent years Hicks has had some second thoughts about this interpretation does not change the basic fact that in 1937 he advanced this interpretation and that Keynes had no objections to it.

The Age of Schumpeter

By HERBERT GIERSCH*

The centenary of Schumpeter's birth coincides with a revival of Schumpeterian economics. Could the third quarter of this century justly be called the "age of Keynes" (John R. Hicks, 1974), the present fourth quarter has a fair chance of becoming the age of Schumpeter. Before giving some substance to this proposition, I shall present a short introduction to Schumpeter's life, work and paradigm.

I

Schumpeter was born in a small place in Moravia, the only child of an Austrian couple. When his father, a cloth manufacturer, died four years later, his mother moved to Graz (Austria) where he attended elementary school until the age of ten. Then his mother married a retired general. For Schumpeter this meant access to Austria's foremost school where he passed with flying colors. At Vienna University (1901–06) he was inspired by Böhm-Bawerk and Wieser, Carl Menger's students. After taking his doctorate in 1906 he spent the summer term in Berlin, was a research student at the London School of Economics and accepted a position at the International Court in Cairo from where he returned to Vienna to submit his habilitation thesis. Shortly afterwards (1909) he became associate professor in Czernowitz (now in the Soviet Union) and, two years later (1911), full professor in Graz, where he taught until 1919, except for 1913–14. During that year he was visiting professor at Columbia University which gave him an honorary doctor's degree at the age of 31. His last six years in Austria (1919–25) were devoted to nonacademic ambitions which he could not realize, neither as an Austrian Minister of Finance for less than eight months in 1919, nor as the head of a private bank which eventually collapsed in 1927, leaving him with a high personal debt to be paid off. It was with great relief that he received offers from two universities in Japan and Germany, accepting the one from Bonn where he was Professor of Public Finance for seven years. Shortly before Hitler came to power, Schumpeter went to Harvard. He was a cofounder of the Econometric Society, served as its president from 1937 to 1941, elected to the office of president of the American Economic Association (1948), and was designated to be the first president of the newly founded International Economic Association. In January 1950, Schumpeter died in his home in Taconic, Connecticut.

II

Schumpeter's main work as a scholar has three strands: an evaluation of past and current economic theory, starting with his postdoctoral book on the state of economic theory (1908) and ending with the posthumous *History of Economic Analysis* (1954); the elaboration of a theory of economic evolution, starting with *The Theory of Economic Development* (1912) and culminating in *Business Cycles* (1939); and the advancement of a theory of social and institutional change, starting with *Crisis of the Tax State* (1918), culminating in *Capitalism, Socialism and Democracy* (1942), and ending with a paper at the AEA meetings, "The March into Socialism" (1950).

III

Obituaries and later biographic essays[1] allow offering a stylized picture of Schumpeter's fascinating personality.[2]

Schumpeter was highly sensitive to aesthetic values. He always remained the aristocratic gentleman of the late Austrian Empire

*President, Institut für Weltwirtschaft, Düsternbrooker Weg 120, D–2300 Kiel. I am grateful to Karl-Heinz Paqué for valuable comments on earlier drafts.

[1] See the seminal contribution by Gottfried Haberler (1950) and the remarkable paper by Christian Seidl (1982) who succeeds in discarding some old myths about Schumpeter through careful analysis of the historical evidence.

[2] For a more detailed analysis along the following lines, see Karl-Heinz Paqué (1983).

who loved elegant clothing, refined meals, polished manners, cultivated conversations and, above all, beautiful women. His style of writing was baroque, with frequent excursions into seductive side issues, occasionally ending up in mere l'art pour l'art. Even as an economist he seems to reveal an aesthetic bias: in his admiration for Walras and in his enthusiasm for the art of formalizing complex phenomena, an art which lay beyond his own reach.

Schumpeter was a staunch individualist. He loved to "épater les bourgeois," that is, to express shocking minority views even at the risk of isolating himself from the mainstream of political and economic thinking. Always ready to display a wide range of sparkling ideas he excited a large number of students who later became famous economists—Samuelson, Schneider, Smithies, Stackelberg, Stolper and Sweezy to mention only those whose names begin with S like Schumpeter. However, his impact was inspiration rather than indoctrination: in Schumpeter's Socratic view of scholarship, there was no legitimate place for the missionary zeal and the fighting spirit of intellectual sectarianism. Furthermore, his work was too original to permit easy paradigmatic simplification: he advanced into dynamics when mainstream economics was grappling with static optimality; he stuck to microeconomics when the tide of Keynesian macro theory supplied a new generation of economists with a fertile intellectual playground; and he turned to historical methods when econometrics—under Schumpeter's own intellectual sponsorship—began to swamp economics.

In accordance with his social background Schumpeter was inclined to see the world from an elitarian perspective. He regarded clusters of talented people as the driving force behind economic and political history: entrepreneurs who push forward society's technological frontier; a nobility to protect the capitalist system by performing the political functions which are alien to the commercial outlook of the bourgeoisie; and the intellectuals who help to destroy capitalism by undermining its ethical basis in an almost tragic process of critical subversion. Even Schumpeter's unfortunate decision to enter politics in 1919 seems to be in accordance with the role which he saw for himself as a member of the old élite in a period of transition.[3]

IV

The essentials of Schumpeter's thought can best be inferred from his relation to Keynes and the economics of Keynes. Apart from a streak of jealousy which may have distorted his judgement, Schumpeter's apparent dislike of Keynes' gospel had deep roots in basic differences of a paradigmatic nature. Consider his penetrating critique of the *General Theory* which focusses on four crucial points. First, he objected to what he called "Keynes' practice of offering, in the garb of general scientific truth, advice which...carries meaning only with reference to the practical exigencies of the unique historical situation of a given time and country" (1936, p. 791), namely England in the 1930's, a practice which Schumpeter—then a detached observer of worldly events—regarded as appropriate for a politician, but not for a "scientific" economist. Second, Schumpeter objected to Keynes' lighthearted use of economic aggregates, most of all "the extension of the Marshallian cross" (1936, p. 793) to aggregate demand and supply functions, a procedure which the microeconomist Schumpeter deemed to be highly suspect. Third, Schumpeter criticized Keynes' assumption of a given technology with a lack of investment opportunities which appeared absurd to the man who had declared the dynamics of technology, the process of creative destruction, as the very essence of the capitalist system. And finally, there was Keynes' message that unemployment could be attributed to underconsumption and hence to private thrift and an unequal income distribution, a message which, according to Schumpeter, enabled the disciples to destroy "the last pillar of the bourgeois argument" (1951, p. 289). To Schumpeter, the historian of intellectual and institutional change, this message made up the essence of the Keynesian revolution.

[3] Of course, personal ambitions played their part as well. On the whole issue, see Seidl, p. 38.

Behind this fundamental critique we find a social vision which in some crucial respects is diametrically opposed to that of Keynes. Schumpeter's vision takes shape when we recognize how he characterized his great contemporary in a later essay: "He was surprisingly insular, even in philosophy, but nowhere so much as in economics" (1951, p. 274). "He was not the sort of man who would bend the full force of his mind to the individual problems of coal, textiles, steel, shipbuilding. Least of all was he the man to preach regenerative creeds" (pp. 274 ff). This point about "regenerative creeds"—made in 1946—highlights Schumpeter's postwar optimism. The point is gaining more and more relevance in our present phase of slow world economic growth, a phase with cumulating pains of delayed adjustment. In such a phase the faith in the regenerative forces of a decentralized market system has once more become critical for the choice of the appropriate socioeconomic paradigm. Let me take this presumption as a justification for considering now a possible—non-Keynesian—paradigm along Schumpeterian lines of thought, hoping that it may help us to better interpret the present quarter century following the "age of Keynes."

V

Such a post-Schumpeterian paradigm may be stylized in the form of ten basic postulates:

1) The approach is micro rather than macro, socioeconomic (if not socio-ecological) rather than mechanistic. In the spirit of Schumpeter's "methodological individualism" it concentrates on processes rather than outcomes, on voluntarism rather than determinism. Being addressed to current world economic development, it stresses relevance rather than rigor, movement rather than static optimality.

2) Steady-state equilibria may be attractive aesthetic devices, but economic life and history show cycles and discontinuities as a normal feature: sunspot cycles, life cycles, product cycles, election cycles, fashion cycles, seasonal cycles, business cycles, growth cycles, technological revolutions and all sorts of lagged adjustments and overreactions to unanticipated events in the markets for factors and products, for assets and monies. With an unknown future, civilizations can only learn by trial and error; equilibria can only be identified by passing them from the other side, just as the pendulum finds its point of rest only in a process of damped oscillations.

3) What matters most in present circumstances are the driving forces of economic development in advanced countries. Emphasis, therefore, is on the growth and dissemination of knowledge, on pathbreaking entrepreneurs who create new markets and successful "intra-preneurs" who rejuvenate old firms, on credit creation for the supply of venture capital, and on Schumpeterian competition (i.e., on innovative monopolistic competition rather than sterile perfect competition, on oligopolistic rivalry rather than collusive equilibria, on aggressive trading rather than mere arbitrage transactions).

4) In the international economy which Schumpeter mostly neglected—despite an occasional sympathy for U.S. protectionism in what Schumpeter called a "mercantilist, nationalist, bellicose world" (1940, p. 7)—the emphasis for the advanced countries is on free trade rather than fair trade (trade minus competition); for the less advanced countries it is on offensive export orientation rather than protective import substitution, and for North-South relations it is on product cycle goods, private resource transfer and catching-up processes.

5) Elasticities, and notably adjustments involving the supply side, are primarily a function of time because of institutional and technical rigidities and inflexibilities in behavior patterns. The relevant time span is longer than the Keynesian short run (which Schumpeter equated with a forty-month cycle), but shorter than the Marxian long run (which includes the eventual breakdown of the system). In terms of calendar time, we may estimate this medium run to cover two to three decades, so as to include at least one turning point of a Kondratieff cycle in Schumpeter's three-cycle hypothesis.

6) In such a cyclical setting, and with an unlimited potential for the growth of knowl-

edge, stagnation can be taken as a temporary phenomenon unless the economy is overregulated. Even in the absence of new technological revolutions, stagnation will last only until relative prices of factors and goods have sufficiently adjusted to restore the incentive structure: profits and profit expectations must be high enough to induce entrepreneurs to overcome barriers to entry erected in favor of existing suppliers.

7) The real rate of interest may be zero in the model of a stationary state, as the young Schumpeter asserted; in a dynamic world it can turn out to be negative as in the recent phase of unanticipated inflation but will thereafter be correspondingly higher as it is now in the subsequent period of correction, when (*i*) monetary disinflation is not fully anticipated, (*ii*) saving habits in the private sector and spending habits in the public sector are slow to adjust to an increasing demand for loanable funds, (*iii*) investors are slow to shift from excessive capital-deepening to more capital-saving technologies, or (*iv*) investment is clouded with too much uncertainty due to a reorientation in the development process.

8) Uncertainty and limits to growth also result from political attempts at impairing property rights and, if intellectuals are the propelling force, from their influence on the social atmosphere, including the public's attitude towards technical progress, entrepreneurship, self-help, and Schumpeterian competition on a national and international plane. The "march into socialism," however, is not inevitable, as intellectuals in their monopolistic competition are also innovative in producing and propagating alternative models of society or even learn from experience as they often do when they enter practical life or when they live under real socialism.

9) In an open world economy, Schumpeterian competition also prevails among governments and central banks. Such policy competition—as competition elsewhere—is efficient in the medium run as a process of discovery and learning although—or because—it offers unpleasant short-run lessons to the misbehaving countries and central banks. In a (Keynesian) short-term paradigm—so close to the heart of politicians in office—these lessons are denounced as beggar-thy-neighbor policies, thus yielding popular arguments in favor of policy cartels called "coordination."

10) Entrepreneurial talent is in almost unlimited supply, but in some countries it finds productive outlets only abroad, or less productive (or even counterproductive) use in politics and government, in public and private bureaucracies or in the military.

VI

A post-Schumpeterian paradigm has to cover the whole world economy with all its diversity. In accordance with the strength of the (re)generative forces we may distinguish

1) "Advanced Schumpeterian areas" which have plenty of innovating firms and people to act as growth locomotives (for example, parts of the United States and Japan);

2) "Less advanced Schumpeterian areas" which are populated by firms and people who as imitators are active absorbers of foreign technologies and capital (for example, Taiwan, Singapore, South Korea, Hong Kong);

3) "Advanced Keynesian areas" which suffer from distorted factor prices depressing the marginal efficiency of capital and from institutional rigidities impeding the entry of new entrepreneurship so that government deficits and foreign demand are needed as substitutes for autonomous investment (for example, large parts of continental Europe);

4) "Less developed Keynesian economies" which for similar reasons rely on import substitution strategies, government deficits financed by inflation, and hopes for a "New International Economic Order" (for example, Latin America and parts of Southern Europe).

This typology is, of course, not complete; we may further identify "Ricardian economies" which exploit their natural resources and convert them into consumption or other forms of wealth, "Malthusian regions" which find themselves in the population trap, and "Marxian countries" which conduct central planning and state trading.

VII

The geographic base of this post-Schumpeterian paradigm can be systematized by making use of a theory of location derived from the writings of a German economist who must be mentioned today together with Keynes, Marx, and Schumpeter, as he was born 200 years ago: Johann Heinrich von Thünen. In a book published in 1826 Thünen not only prediscovered marginalism (which earned him high praise from Schumpeter) but also developed a center-periphery model for the spatial division of labor on a homogeneous (i.e., non-Ricardian) plane surrounded by a wilderness. Thünen took the central market as given, but the center can well be explained by (*i*) the provision of a public good called law enforcement or defense which is—as Adam Smith has taught us—a prerequisite for the division of labor, or by (*ii*) assuming a point of superior resource endowment with high quality land, raw material deposits, or favorable climatic conditions which yield a Ricardian rent, or by (*iii*) introducing external economies of agglomeration which generate knowledge to be used by entrepreneurs and which, therefore, produce the Schumpeterian transitory rent which we call profit. In the real world we can depict many centers and a hierarchical order of them, but the major center-periphery systems in the world economy have turned out to be supranational like the Pax Britannica of the nineteenth century, the Pax Americana of the twentieth century or the present triple center system of North America, Western Europe, and Japan (leaving aside the Marxian center in Eastern Europe). As economic development essentially consists of exploiting knowledge, a social atmosphere conducive to knowledge production must be taken to be the most important element in the formation of growth centers. This is why MIT and Stanford have become the Mecca and Medina of achievement oriented thinkers and operators; why some countries like Japan and France strive hard on the technology front; and why large parts of Europe where equality was considered to be more important than (Schumpeterian) excellence presently tend to fall behind in world economic development.

VIII

The present quarter of the twentieth century is likely to become Schumpeter's age, since autonomous investment—at least in Europe—has become so weak during the last decade that the sociopolitical focus is shifting towards regenerative forces which seem to have been weakened by extensive reliance on monetary-fiscal management. The medicine of boosting demand surely helped in the short run. Where it was periodically withdrawn for the sake of fighting inflation, it even helped over a number of business cycles. In the medium run, however, it was bound to weaken the patient's motivations and his overall physical strength. This is so because any kind of unconditional support to suppliers—from full-employment guarantees, fine-tuning promises, and programs of industrial policy right down to specific subsidies and sophisticated protective devices against import competition—must be presumed to produce a dependence effect and gradually weaken the need to adjust, and with it the need that is proverbially considered to be the mother of invention and innovation. In more general terms, it can be said that permissive policies promoted the march into the soft society which, for lack of a hard constitutional dividing line between social goals and individual responsibilities, became overwhelmed by populist pressures. Moreover, permissive policies offer incentives for rent seeking, thus distracting entrepreneurial talent from future-orientated activities to lobbying and distributional issues. Eventually, governments find themselves at the limits of the tax state which the young Schumpeter clearly foresaw.

The post-Schumpeterian paradigm proposed here includes the vision of a turnaround to be brought about by regenerative forces. Where can they be found? 1) We observe disillusionment with government policies, including the welfare state, and an increasing sensitivity to fiscal issues. 2) We witness the growth of the underground economy which has a good chance of becoming a school for entrepreneurship, similar to the black market in Europe's initial postwar phase before the miraculous reconstruction, and also a spectacular growth of self employment and job creation in new firms for new products in some parts of Europe as well as in the United States. 3) We visualize how severe lapses from full employment are about to weaken rigid labor market institutions even in syndicalist Europe where a tendency towards greater balkanization and flexibility has developed. 4) And we take it that further progress in telecommunication will not only boost investment by itself but also by facilitating decentralized decision making. Should these new technologies promote decentralized production, they can be expected to further improve the incentive structure by making the old factory system obsolete and with it the rigid labor market institutions inherited from the past.

The turnaround may be firmly expected but it can come about only gradually. At least in Europe, dynamic forces are hampered by encrusted institutions. Perhaps technical progress alone will suffice to overcome institutional obstacles by carrying innovative activities into unregulated fields. But in many regions and industries on both sides of the Atlantic, a temporary crisis may be both inevitable and necessary to bring about the destruction which Schumpeter considered to precede creation or to go along with it. By widening the spread in earnings between forward-looking and backward-looking persons, firms, industries, and countries, the turnaround will strain widespread feelings for equity and the so-called social-democratic concensus. And there will be no reward for tolerating inequality until the turnaround has actually led to faster growth. Rawlsians will, therefore, have to stretch their implicit time horizon beyond the Keynesian short run, so as to include the medium run which is the time horizon required for starting and successfully completing adjustment processes on the supply side. Hence time will remain a resource in short supply, but in high demand.

As an indicator of how much the time pattern of preferences diverges from the time pattern of opportunities and necessities, I submit taking the dramatic change from excessively low to excessively high real rates of interest in the world economy. In my view this change reveals how much society in the past has allowed itself to live at the expense of its future. Lower real interest rates will eventually come back, albeit not by decree or a different monetary regime, but only after the world has again learned to pay its tribute to the laws of efficiency for the benefit of capital formation.

In the international context the turnaround will not get underway before a Schumpeterian perspective has gained widespread support in the industrialized and newly industrialized countries of the North. Only after more northern entrepreneurs, firms, and governments have adopted forward-looking strategies that anticipate the changing international pattern of comparative advantages, will southern entrepreneurs, firms, and governments feel encouraged to link themselves more closely to the northern growth locomotives. When this has happened, an accelerated world dynamics will raise the marginal efficiency of capital in the South, and thereby will—in a virtuous circle—promote a sustainable private resource transfer to the South and also diminish the high level of uncertainty presently prevailing on international capital markets.

Once world economic growth has reaccelerated—say towards the end of this decade or in the 1990's—Schumpeter in his Valhalla can step back from the intellectual leadership which this essay attributes to him in the succession of Keynes. But for today the question is whether the man who wanted to be the greatest economist of his time could be imagined to agree with the preceding attempt at bringing his version in line with the course of economic history after his death. A competent answer must be reserved to those who were lucky enough to know him personally. So I have to be quiet. What remains is the wider question whether any Schumpeter-based paradigm has relevance at all for this quarter century, but here the judge can only be future history itself.

REFERENCES

Haberler, Gottfried, "Joseph Alois Schumpeter 1883–1950," *Quarterly Journal of Economics*, August 1950, *64*, 333–72.

Hicks, John R., *The Crisis in Keynesian Economics*, Oxford 1974.

Paqué, Karl-Heinz, "Einige Bemerkungen zur Persönlichkeit Joseph Alois Schumpeters," Kieler Arbeitspapier No. 193, Institut für Weltwirtschaft, Kiel, December 1983.

Schumpeter, Joseph A., *Das Wesen und der Hauptinhalt der theoretischen Nationaloekonomie*, Leipzig 1908.

______, *Theorie der Wirtschaftlichen Entwicklung*, Leipzig 1912 (English: *The Theory of Economic Development: An Inquiry into Profits, Capital, Credit Interest, and the Business Cycle*), Cambridge, MA, 1934.

______, *Die Krise des Steuerstaats*, Leipzig, 1918.

______, book review of Keynes, J. M., *The General Theory of Employment, Interest and Money, Journal of the American Statistical Association*, 1936, *31*, 791–95.

______, *Business Cycles*, Vols. I, II, New York; London, 1939.

______, "The Influence of Protective Tariffs on the Industrial Development of the United States," an address before the Academy of Political Science, April 11, 1940.

______, *Capitalism, Socialism and Democracy*, New York, 1942.

______, "The March into Socialism," *American Economic Review Proceedings*, May 1950, *40*, 446–56.

______, "John Maynard Keynes 1883–1946," *American Economic Review*, September 1946, *36*, 495–518; reprinted in *Ten Great Economists, From Marx to Keynes*, London 1951, 260–91.

______, *History of Economic Analysis*, Vol. I, II (from manuscript by Elisabeth Boody-Schumpeter), New York 1954.

Seidl, Christian, "Joseph Alois Schumpeter in Graz," Research Memorandum No. 8201, Universität Graz, August, 1982.

von Thünen, Johann Heinrich, *Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*, Hamburg, 1826.

Marx's Contributions and their Relevance Today

By JOHN G. GURLEY*

I shall speak first of Marx's major contributions and then whether, if he returned today, he would see any close relation between them and the present-day world.

I. Marx's Major Contributions

Karl Marx made at least seven major contributions to political economy.

First, he established a framework—the materialist conception of history—for analyzing economic, social, and political changes over long periods of time. Marx showed that a society's social or class relations ultimately became impediments to the further development of its productive forces. In order for the productive forces to gain the conditions for their further advance, the rising class associated with the economic expansion would have to overcome, in one way or another, the prevailing ruling classes who were tied to the older productive forces. Once the relations of production were radically altered, political, ideological, and cultural changes would follow. Marx concluded that all class societies, including capitalism, are transitory.

Second, Marx investigated the production and circulation processes of industrial capitalism, from which he formulated a labor theory of value for analyzing the exploitation of workers by the capital-owning class. In this analysis, Marx found the origin of surplus value, the methods employed by capitalists to increase surplus value, and the role of the price system in redistributing the surplus value among capitalists. Marx concluded that the working class was bound to suffer impoverishment relative to the growing wealth around it, and, at times, absolutely.

Third, Marx studied the processes of capital accumulation—that is, of investment, growth, and cycles—in capitalist societies. He showed that, during the accumulation process, there was a strong tendency for the rate of profit to fall and hence for the eventual retardation of capital accumulation. The ensuing recession restored the conditions required for another upswing, including the replenishment of the industrial reserve army of the unemployed and the strengthening of capital through mergers (centralization of capitals) and write-offs of redundant capital goods. Over the long run, according to Marx, the capital accumulation process created both wealth and poverty, it both drained and refilled the army of the unemployed, and it spawned both increasingly larger enterprises and a proliferation of small ones—the former comprising the monopoly sector and the latter a crowded and intensely competitive sector of small capitals. One of Marx's conclusions from these investigations was that periodic business cycles were endemic to capitalism, as natural and as inevitable as changes of the season; there was no remedy for them within the system of capitalism itself. The cycles were also functional for the (temporary) survival of the system.

Fourth, one can find an *economic* theory of the state in Marx's writings. His researches led him to suppose, as I have just noted, that the state could not alleviate the commercial crises of capitalism, nor even the monetary panics which were but a phase of the broader crises. He was especially skeptical about financial policies as cures for what he took to be decennial cycles, although he did believe that certain budget and bank measures had temporary effects on economic activity—for example, easy money policies, he said, could "keep the shopkeeping world in a good mood," presumably until the next crisis. On the other hand, the state could effectively intervene with economic legislation to support the capitalist class against any growing strength of workers, or to prevent the ruling class from destroying, through excessive

*Stanford University, Stanford, CA 94305. I am indebted to Kenneth Arrow for valuable comments on an earlier draft.

hours and worsening conditions imposed on workers, the labor power on which the existence of the capitalist system itself depended. Marx's conclusion was that, while the state could prevent excesses, it could not alleviate the panics and cycles of the capitalist system.

Fifth, Marx explained how workers are mystified by the system of capitalism, alienated within its production sphere, and misled by false solutions to their problems. Marx alleged that capitalism presents itself on the economic surface in distorted forms—that what it is really like, within its deep recesses, is quite different from what its facial expressions suggest. These exterior distortions, Marx maintained, lead to illusions about capitalism, and the illusions in turn are used by bourgeois ideologists to mystify the workings of the system. Thus, a web of mystification is spun around this mode of production, hindering a clear understanding of its true nature. Marx also explained how workers become alienated in their work, confronted by the outside power of capital, slaves to natural necessity, subject to the interfering power of other men, and unable to exercise fully their own essential powers. As alienated beings, workers lose their integrity, freedom, and understanding of the world around them. Finally, Marx claimed that the working class was repeatedly deceived by numerous anarchist, reformist, and bogus socialist movements that drew it away from scientific socialism, consuming its time and energies into deadends of false promises.

Sixth, Marx investigated the future course of global capitalist and socialist development, and he examined the impact of capitalist expansion on less-developed countries. His main conclusions, although modified from time to time, were that 1) capitalism would spread from Europe to the rest of the world, 2) the impact of capitalist expansion would be both brutal and progressive, 3) the first proletarian revolutions would occur in the most advanced capitalist countries, and 4) the rest of the world would eventually become socialist, too.

Marx believed that the capitalist mode of production contained an inherent need for ever-expanding markets for its money and commodity capital, and that this outward movement was furthered by the tendency for the rate of profit to fall within the advanced countries. In the resulting process of expansion, the European bourgeoisie compelled "all nations, on pain of extinction, to adopt the bourgeois mode of production," including those countries that could change only through the shock of European colonial aggression. Capitalist expansion abroad was both necessary and savage, both constructive and destructive.

Marx also studied the many connections between revolutionary ferment in the colonial world and proletarian revolutions in the West. He investigated the possibilities of bourgeois revolutions and independence movements in the less-developed areas, particularly in those countries richer in feudal vestiges than in capitalist institutions and practices, concluding that the army would frequently play the star role in nationalist movements and local insurrections.

Seventh, Marx sketched outlines of the future socialist and communist societies. He anticipated that society, after the revolution and overthrow of capitalism, would go through a transition period in which the state would take the form of the dictatorship of the proletariat, the means of production would be in the hands of the working class, planning would replace the anarchy of markets, the productive forces would be greatly enlarged, and a momentum toward full communism would be established. When full communism was attained, the productive forces would be ample; class institutions, including the state and party, would disappear along with social classes; commodities and money would give way to products that were produced and distributed according to an overall plan, the distribution principle being to fulfill the needs of people; and society would follow the ruling principle of "the full and free development of every individual." Marx also anticipated that a socialist-communist world would see the decline and disappearance of nationalism and supernaturalism, and the advent of a true democracy that transcended classes. Beyond the "realm of necessity," Marx concluded, lay the true "realm of freedom," in which labor is no longer ruled by an external compulsion but instead becomes an end in itself, the development of human energy leading to fully matured individuals who give free scope to their natural and acquired powers.

II. Marx's Return and Assessment

Suppose that Marx were to rise from the dead and survey our world of theory and practice. What would he say about the relevance of his seven major political-economy contributions to the contemporary scene?

First, Marx would see confirmation for his conclusion that capitalism is transitory. At the very time of Marx's death, the leading capitalist countries began a final wave of colonization, and by 1914 they had completed their domination of most of the globe. Since that time, the tide has run the other way, as Marxism has conquered one area after another, until it now comprises about one-third of the earth's land and population. Marx would see much evidence that revolutionary socialism is replacing capitalism around the world.

However, he would find it quite difficult to explain this fact with the materialist conception of history, as he left it. This is because Marxian socialism has replaced only immature capitalist or even pre-capitalist societies, while it has made few inroads in the most mature capitalist countries, where, according to his theory, the first proletarian revolutions should have occurred. Socialism based on poverty and the immature development of both the proletariat and the bourgeoisie would present a powerful challenge to Marx.

He would also discover that, in the most advanced capitalist countries, the proletariat had gained in many ways, despite the continuing exploitation against it, and that new middle classes had arisen, contrary to his expectation of class polarity. Capitalism based on the growing affluence of the working class and on the seemingly endless process of the maturation of this mode of production would also present a puzzle to Marx.

Moreover, Marx would find that his materialist conception of history had not fared well among economists generally. He would see that much of economics continues to be dominated by a type of philosophic idealism, similar to the outlook that he and Engels railed against in *The German Ideology* —that is, the attitude that ideas do not come from the real world, but they are, nevertheless, the principal force changing it. He would observe that many economists were not attracted to his opposing view that the real world is the source of our ideas, and that the principal force changing the real world is not ideas but class struggles, which themselves grow out of the development of society's productive forces. Marx would, of course, contend that these economists favor idealism because it serves the interests of the bourgeoisie by diverting attention from the real world of class antagonisms and directing it to a realm in which ideas seem to originate by feeding on other ideas, in an endless procession, and in which the relevant struggles are not between contending classes but between contending ideas. He would conclude that, in this regard, little had changed in the past century.

Second, if Marx returned today, he would certainly believe that workers were continuing to be exploited within the production sphere, where substantial shares of output continued to be claimed by the owners of capital goods, as though sheer ownership was productive. He would see here essentially the same structure that he had left—one in which a capitalist class owned the major means of production, controlled the production process, claimed ownership of the commodities produced by labor, and realized and appropriated surplus value through the sale of the commodities.

At the same time, Marx would note the almost total disappearance among economists generally of the labor theory of value and of his analysis of exploitation. He would not miss the fact that bourgeois economic theory, of the present day as in his own day, continued to emphasize gradual changes, not revolutionary leaps; maximizing individuals, not warring classes; consumers and producers, not workers and capitalists; general equilibrium and harmony, not imbalances and disharmony; relative scarcity as the human condition, not brutal exploitation; individual rationality and order, not mystification and anarchy. In all of this and more, Marx would see the persistent attempt by bourgeois economists to hide the true production relations of capitalist society behind a facade of superficial market phenomena; he would see not the invalidity of his labor theory but a vindication of his views about the links between ruling classes and ruling ideas.

Third, a revived Mark would learn that most of what he wrote about the capital accumulation process was still valid. The capitalist system, he would find, has continued to generate business cycles, despite Keynesian policies, fine-tuning, and automatic pilots. Further, Marx would perceive that the cycles appear to many analysts today, as they did to him, as necessary for the survival of the system—whether these analysts speak directly in such terms or only indirectly in terms of slowing the upturn to prevent a new round of inflation, to moderate wage demands, or to fashion a sustainable recovery. Marx would also conclude that capitalism remains dependent on a reserve army of labor that acts to modify wage and other demands of the employed wage workers. In this connection, he would notice the 34 million unemployed today throughout the industrial capitalist world, and he would certainly be interested in the massive supplies of cheap labor from the underdeveloped areas of the world that have supported capitalist prosperity during the years since his death and particularly in most recent times. He would also see that, during his centennial absence, labor forces in advanced capitalism had been steadily proletarianized, according to his expectation, as the self-employed had declined sharply and wage labor had increased. However, as I previously noted, Marx would have to account for the greater complexity of class structures than he had anticipated.

Fourth, Marx would probably be surprised by the extent to which the state has gone to try to stabilize the economy; by the much larger roles played by government expenditures, revenues, and money in these efforts; and by the heightened ability of financial institutions to weather financial storms. He would undoubtedly attribute the changes he observed to the system's responses to the continuing severity and frequency of business cycles, which he was one of the first to analyze.

Marx would learn little or nothing that would shake his belief in the inherent instability of the capitalist system and in the inability of government policies to do much about it. He would consider the faith in Keynesianism by some economists and in *laissez-faire* by others as equally naive. He would not expect the state, even if it could, to eliminate the cycles that capitalists so urgently need to discipline the work force and to correct other imbalances of the boom period.

Fifth, in Marx's day, capitalism produced not only illusions and alienation among workers, but a variety of anarchist, socialist, and reformist doctrines and movements that competed successfully against Marxism. Marx was involved much of his life in trying to overcome these barriers to the workers' correct understanding of their problems under capitalism and the solutions to them. He believed that workers were led astray partly because they felt unable, in their state of alienation, to utilize their inherent powers in the right directions, partly because they were misled by the surface manifestations of capitalism and so failed to discern the exploitative class relations that were basically responsible for their miseries, and partly because their energies and time were misdirected down blind alleys by false doctrines that were produced by the particular way capitalism developed. If Marx came back today, it is likely that he would see all of these phenomena still at work to thwart proletarian revolutions in the advanced capitalist countries.

He would be dismayed by the great and rising strength of nationalism, by the continuing debilitating influence of religion on the working class, and by the bait of bourgeois democracy that workers still snapped up.

After observing the extent of alienation, mystification, and superstition on all sides of him, he would be most interested in the prevailing assumptions of rationality within economic theory. In this connection, he might be told about Keynes' attribution of mainly animal spirits to risktakers, Schumpeter's deep distrust of the assumption of close rational calculation by entrepreneurs, and Veblen's suspicions about consumer sovereignty and rational choice. No doubt Freud would be discussed, too. In the end, Marx would see no reason to alter his impression that the capitalist world contains many confused and misinformed people. But he would wonder, in thinking back on his own analysis, why the continued growth of the productive forces and the presumed consequent growth of workers' understanding of the world around them had not done more to mitigate their irrational strains.

Sixth, to regard to Marx's investigation of the future course of global capitalist and socialist development and to his examination of the impact of capitalist expansion on less-developed countries, he would be surprised at how wrong his first conclusion was. Instead of capitalism spreading and developing in European fashion throughout the non-European world, much of these areas largely skipped the capitalist stage, jumping from immature capitalism or pre-capitalist societies directly into socialism, and most of the rest have been following different capitalist paths than that taken by the advanced capitalist nations of Marx's day. To understand all of this, Marx would have to know what Lenin knew about the later worldwide development of capitalism plus still later facts about how the wealthy countries, in various ways, blocked the economic and social development of the poor ones, about the internecine wars among the capitalist powers for the redivision of the world, about the growing strength of nationalism as an ideology, and about the increasingly rich experiences of the many Marxist-Leninist movements that built on the Russian and Chinese revolutions to gain similar ends elsewhere. Marx would have much to learn and to ponder in this area.

His contemplations would easily—or uneasily—extend to his third conclusion— that the first proletarian revolutions would occur in the most advanced capitalist countries. In this, he would have to conclude that he was also dead wrong. Once again, to understand what had happened, Marx would have to know what Lenin knew—specifically, in this case, that the forceful spread of colonialism around the world provided ample room for the further development of capitalism's productive forces, which produced the super surplus value to bribe key strata of the working classes in the advanced countries, and at the same time provided higher living standards to many others. Marx would also have to know that the struggles of the workers won them many concessions from the capitalists and the state, and that the expansion of the productive forces created an increasing variety of conservative middle-class occupations, calling for professional, technical, administrative, and scientific talents and training. Society, therefore, was not split, as Marx predicted, into two hostile classes with nothing in between, but instead there evolved a much more complex class structure.

His second conclusion that the impact of capitalist expansion would be both brutal and progressive he would consider to be true, while his fourth conclusion that the rest of the world would become socialist only after the most mature capitalist countries had reached that mode of production—that reasoning was simply a corollary of the first and third conclusions and so equally in error.

In considering the developments since his death in both the advanced and less-developed areas of the world, Marx would be compelled to reassess his materialist conception of history, which led him to erroneous conclusions. His theory of history would have to allow for the growth of capitalist productive forces on a global scale, for changes within the bourgeois relations of production and exchange that would accommodate and not stifle the expanding productive forces, and for the deeper penetration of bourgeois ideas within the working classes.

Seventh, Marx's sketch of socialism and communism was based on the ongoing development and transformation of advanced capitalist societies, in which the productive forces became ample and relations of production had become ripe for their overturn —that is, a highly class-conscious proletariat; a capitalist class left largely without function; the disappearance of the peasantry, the urban petty bourgeoisie, and all other classes associated with previous modes of production; the readiness of the Party to assume its socialist functions; and planning already in practice within large enterprises.

If Marx returned today, he would not likely recognize "his" socialist society anywhere in the world. He could hardly imagine socialism on the basis of material poverty, in which peasants comprised most of the population and agriculture was the principal economic pursuit. He would not recognize "his" socialism where the working class was largely dispossessed of power; where there was little or no momentum toward full communism, not to speak of the realm of freedom lying beyond the realm of necessity. He would not see "socialism" in any of the advanced industrial economies, for he would quickly become aware that throughout these societies political power was in the hands of the capitalist classes; that markets, commodities, and money were more important than ever; that proletarian internationalism did not exist; that workers remained bent under the weight of bourgeois propaganda and were heavily under the sway of nationalism, supernaturalism, and other demeaning passions.

I think that it would be obvious to Marx, when he had surveyed the world, that much of capitalism's and socialism's difficulties could be explained by the absence of strong, leading classes in those societies. He would see that the advanced capitalist economies are suffering because of the inadequate and declining strength of their capitalist classes. The socialist countries are foundering because their intended leading class—the working class—is generally impotent before a bureaucratic state and Party. And in the less-developed countries, development is retarded by the lack of ruling social classes to chart the way; leaving the door open for the military or civilian bureaucrats. Marx would not find it too arduous, within his own analytical framework, to account for these historical results. And while undertaking that task, you can bet your life that he would begin organizing the proletariat.

The Limits of Neoclassical Theory in Management Education

By DAVID J. TEECE AND SIDNEY G. WINTER*

The productivity slowdown and the competitive difficulties experienced by many American firms in domestic and international markets have triggered considerable research into the reasons for the observed changes. Although rarely featured in systematic analysis of productivity trends, various writers have suggested that management miseducation may be part of the problem. In an award-winning article, Robert Hayes and William Abernathy assert

> [R]esponsibility for this competitive listlessness belongs not just to a set of external conditions, but to the attitudes, preoccupations, and practices of American managers.... During the past two decades, American managers have increasingly relied upon principles which prize analytic detachment and methodological elegance over insight, based on experience, into the subtleties and complexities of strategic decisions. [1980, p. 77]

Hayes and Abernathy focus primarily on the overutilization of techniques that deflect attention from long-run technological development to portfolio management, financial control, and the like. Thomas Peters and Robert Waterman (1982) develop this theme further. Their much-acclaimed study of America's excellent companies heaps ridicule on the rational approach to management and asserts that American business schools have a good share of blame for the competitive failures of American industry.

The rational approach to problem solving is, of course, central to economic science. The report by Robert Gordon and James Howell (1959), which defined a watershed in American business education, strongly advocated the adoption of analytical approaches to management education. Economics as a discipline[1] and economists as a profession have played a central role in the implementation of this key recommendation.

While American management practices undoubtedly have shortcomings that are not reflections of corresponding shortcomings in American management education and while management education undoubtedly has limitations that are not attributable to the excessive influence of economics and economists, there is nevertheless a definite correspondence between the symptoms of the patient and the symptoms that an overdose of textbook economics would tend to produce. Moreover, business schools and business school faculties do have a pervasive influence, not only in the campus classroom, but also through consulting, executive education, and the media. The principal purpose of this paper is to identify the particular constituents of contemporary economic thinking that may be implicated in the failures of American management. Of course, it is not the case that attempts at rational analysis are inevitably counterproductive or that economic analysis necessarily distorts managerial decision making. The issues involve matters of balance and emphasis and the role of discipline-based perceptual blinders that make imbalance and misemphasis

*School of Business Administration, 350 Barrows Hall, University of California, Berkeley, CA 94720, and School of Organization and Management, Box 1A, Yale University, New Haven, CT 06520, respectively. We wish to thank Therese Flaherty, John Roberts, and Victor Goldberg for useful comments and criticisms.

[1] We treat modern financial and accounting theory, as well as much of decision sciences/management sciences, as subspecies of economics. They employ many of the same analytical techniques and methodological procedures.

both inevitable and invisible. We will concentrate our attention on features of the discipline that, we believe, usually do have unfortunate consequences for education in business schools.[2]

I. Management Problems and Economic Analysis

Most management problems are ill-structured. They are messy, involving complex interdependencies, multiple goals, and considerable ambiguity. Problems are themselves human creations, and their nature is much dependent on the conceptual lens through which they are viewed (James Griffin and Teece, 1982). Managerial problems are cognitions and recognitions which are products of the manager's conceptual imagination. For this reason, different analysts often can and do conceptualize problems in different ways. It is easy to fall into the trap of working on the wrong problem—what Ian Mitroff (1974, 1977) has referred to as a "type three error."

The problem-solving process includes sensing, defining, modeling, solution derivation, implementation, and the monitoring of results. The processes and procedures that managers employ in solving problems are frequently as complex and disorderly as the problems themselves. They may well be "intendedly rational" (Herbert Simon, 1957), but rarely indeed do they follow an algorithmic path from an obvious or firmly established problem definition to a logically dictated solution. The blending and blurring of conceptually distinct categories such as fact and value, discussion and action, deduction and inference, individual interests and group goals, persuasion and coercion are commonplace. The diversity of the participants implies a corresponding diversity in the rationalities that guide the search for solutions. The economist's special brand of rationality has no special place in the repertoire of problem-solving approaches, whatever claims may be made for it as a universal logic of decision. Indeed, in some contexts economists have long been wont to emphasize that their characterization of rationality does not pretend to describe how decisions are actually taken or ought to be taken. Given that problem definitions and models adopted cannot be perfect in real situations, economic approaches to decision making may well encourage characteristic forms of "type three error." Certainly the empirical proposition that the logic of economic rationality defines a uniquely valuable approach to actual problem solving remains unestablished.

The discipline of economics in general, and formal economic theory in particular, is shaped by a concern with problems that are very different from the management problems just described. To begin with, economists (academic economists particularly) are ordinarily concerned above all that their arguments be found persuasive by other economists. They are relatively sheltered from the fact that there are brands of rationality that compete with their own, and rarely suffer in their professional lives the discomforts and anxieties of reliance on indispensable expertise operating from an alien conceptual framework. As a result, they "respond rather weakly to the ideal of 'seeing the problem whole'" (Richard Nelson and Winter, 1982, p. 405) and are thus ill-equipped to deal with the complexity and diversity of management problems. Second, economics as an empirical science has long been determinedly oblivious to the problems of predicting the behavior of the individual decision unit, and has focused its attention and developed its specialized tools for the statistical analysis of patterns of behavior in whole populations of economic actors. The fact that very different success criteria and information resources are associated with normative study of the problems of the individual entity is often missed and when noted, often underestimated in importance. Finally, and perhaps most importantly, the dominant mode of theorizing in contemporary microeconomics tends to distance the discipline from management problems, with the important exception of problems relating primarily to the function-

[2] We base our judgments partly on our familiarity with the textbooks in price theory and managerial economics that are usually employed in the basic economics courses, and partly on our casual-empirical knowledge of what actually goes on in some schools. Needless to say, we try to avoid the mistakes that we deplore when we teach the subject ourselves. It is not easily done.

ing of organized auction markets operating under high information conditions (i.e., finance). The dominant mode combines unquestioning faith in the rational behavior paradigm as a framework, relative indifference to the delineation of the empirical phenomena that are thought to require theoretical explanation, and a delight in the construction of "parables of mechanism." Such parables provide a sharply defined view of an imaginary world in which the logic of a particular economic mechanism stands out with particular clarity. The insights generated by this method often seem valuable and compelling, but unfortunately there is often no attempt to bridge the vast gulf that separates the simple imaginary world with its isolated mechanism from the complex real world in which some analogous mechanism may, perhaps, operate.

Without doubting the legitimacy or importance of the concerns and objectives that have shaped the discipline of economics, one can doubt very seriously that the discipline thus shaped makes a wholly constructive contribution to management. The following section examines some areas where such doubt seems particularly well justified.

II. Economic Theory and Business Realities

A. *Underemphasis on Dynamics*

Most management issues centrally concern dynamics. Economic theory, on the other hand, deals almost exclusively with static equilibrium analyses. In recent years, much greater attention has been given to theoretical formulations which are dynamic in nature, but formal modeling endeavors of this kind are often exceedingly difficult to perform. Accordingly, only very simple problems can be dealt with mathematically. While comparative statics is one way to get at dynamic issues, it suffers from inattention to the path to equilibrium, a matter which is usually exceedingly important.

B. *Treatment of Know-How*

The production and utilization of technological and organizational knowledge is a central economic activity that is handled in a most cavalier way within economic theory. By far the most common theoretical approach is simply to take technology as given, ignoring entirely the fact that the options open to a manager almost always include an attempt at some degree of innovative improvement in existing ways of doing things. On the occasions when this pattern is broken by explicit attention to technological change, the treatment of states of knowledge and the changes therein is often simplistic and undifferentiated. It is common to assume that technology is uniformly available to all, or, if technology is proprietary, that it is embedded in a "book of blueprints." However, in reality, know-how is commonly not of this form. It is often tacit, in that those practicing a technique can do so with great facility, but they may not be able to transfer the skill to others without demonstration and involvement (Teece, 1981). To assume otherwise often obscures issues relating to the generation and transfer of know-how.

In general, the fact that technological and organizational change is such an important and pervasive aspect of reality, and yet so peripheral in economic theory, may be the single most important consideration limiting the contribution of economics to contemporary management.[3]

C. *Inadequacy of the Theory of the Firm*

With little exaggeration, we can assert that economics is only now beginning to develop a theory of the firm. To be sure, textbooks contain chapter headings labeled "the theory of the firm," but on closer examination one finds a theory of production masquerading as a theory of the firm. Firms are typically represented as production functions, or, in some formulations, production sets. These constructs relate a firm's inputs to its outputs. The firm is a "black box" which transforms the factors of production into outputs, usually just one. Firms are thus single product in their focus. If multiproduct firms exist,

[3] After a recent survey of MBA curriculums at leading business schools, F. M. Scherer concluded that "at maximum, only 12% of MBA students acquired a systematic full-term exposure to the mainstream questions of technological innovation management" (1982, p. 8).

then they are flukes in that they have no distinct efficiency dimensions (see Teece, 1980, 1982).

The boundaries of the firm—the appropriate degree of vertical, lateral, or horizontal integration—thus lie outside the domain of the traditional economic analysis. Moreover, textbook theory is completely silent with respect to the internal structure of the firm. In short, the firm is an entity which barely exists within received neoclassical theory. The only dimension of the firm's activities which is given much attention is the volume of its output and the price at which that output is sold.

D. *Suppression of Entrepreneurship*

Because equilibrium analysis occupies such a dominant position within received theory and because change is so often modeled as a movement from one equilibrium condition to another, the role of entrepreneurship tends to be downplayed, if not outright suppressed. In fact, "it may be said quite categorically that at present there is no established economic theory of the entrepreneur. The subject area has been surrendered by economists to sociologists, psychologists, and political scientists. Indeed, almost all the social sciences have a theory of the entrepreneur, except economics" (Mark Casson, 1982, p. 9). He identifies one villain as the very extreme assumptions about access to information which are implicit in orthodox economics. Simple neoclassical models often assume that everyone has free access to all the information they require for making decisions, thereby reducing decision making to the mechanical application of mathematical rules for optimization. This trivializes decision making, and makes it impossible to analyze the role of entrepreneurs. The development of a theory of entrepreneurship—or at least a theory which does not suppress the importance of entrepreneurship—is of critical importance to economic science as well as to management education.

E. *Stylized Markets*

In neoclassical markets, transactions are performed by faceless economic agents operating in impersonal product or factor markets. While there is some consideration given to the role of reputation effects, the immense variety of institutional supports to market processes—such as trust, friendship, law, and reciprocity—are barely recognized.

Not only are markets characterized by a variety of information conditions, but they differ widely in the frequency with which transactions and the opportunities for and costs of disruption occur (compare the sale of nuclear power plants with the sale of a bushel of wheat). Intermediate markets and relational contracting (Oliver Williamson, 1979) are virtually absent from the textbooks and most advanced theorizing. By stripping out the institutional foundations of market structure, the conventional tools of economic analysis are rendered impotent before many strategic management problems.

III. New Directions

As a result of these distortions, neoclassical economic theory has not brought us very far along the road towards detailed explanation, let alone prediction. All empirical assumptions upon which theories are based are qualitative in character and are quite vague and general. So there are few operational predictions at which the theory arrives. It is rare in positive science to find so elaborate a theoretical structure erected on so narrow and shallow a factual foundation.

Despite these serious shortcomings, there is a basis for optimism. The field of industrial organization has for some time now provided an intellectual framework for dealing with a limited range of strategic management issues (see, for example, Michael Porter, 1980, and Richard Caves, 1984). Equilibrium models of dynamic competition are now also commonplace. Unfortunately, these models focus on the nature of competition with respect to a single variable (for example, irreversible capacity investment, lumpy investments, learning by doing, *R&D*) and are completely unable to deal with interconnectedness. Tacit collusion, reputation effects, and incomplete information problems are also being addressed, but mainly within the structure of very stylized models which turn out to be highly sensitive to initial conditions

(David Kreps and Robert Wilson, 1982; Paul Milgrom and John Roberts, 1982). More promising is the work of Williamson (1981) and others which adopts a bounded rationality framework and takes the transaction as the fundamental unit of analysis. The approach highlights the role of market frictions and of dedicated transaction-specific physical and human capital. A positive theory of the firm's efficient boundaries is emerging, with strong normative implications for management. Progress is also being made in evolutionary modeling of industry dynamics, utilizing a combination of analytical and simulation techniques (Nelson and Winter). Even so, our understanding of industry dynamics is still rudimentary, and the problems identified in Section II remain.

In general, the prospects for a larger and more constructive contribution to management from economics depend on the prospects for three sorts of change in economists' proclivities. There is, first of all, a need for greater diversity and flexibility of theoretical approach, and particularly a greater willingness to trade off some of the aesthetic advantages of simplified models against the virtue of a greater contact with a complex reality. Flexibility is also important with respect to the contributions and criticisms coming from other disciplines that bear on management problems. Second, there is a need for a much stronger base of empirical work, particularly in areas relating directly to management problems, but also in economics generally. Economists' predilection for deductive reasoning needs to be balanced by the discipline of systematic fact finding, as is the case in the natural sciences. If this does not begin in business schools, we despair that it will ever begin. The pursuit of a more fundamental understanding of the firm and production may eventually lead into engineering, as well as into organizational behavior. "True advance can be achieved only through an iterative process in which improved theoretical formulation raises new empirical questions and the answers to these questions, in turn, lead to new theoretical insights" (Wassily Leontief, 1971, p. 5).

In the teaching of economics, there is a need for much greater care in selecting from the impressive range of contemporary economic thinking the principal topics, concepts, and issues that are likely to be useful to future managers. The selection should deemphasize the economists' parable of competitive markets and equilibrium and seek instead to impart understanding of the economic mechanisms that visibly and forcefully shape the managerial decision environment in a rapidly changing world.

REFERENCES

Casson, Mark, *The Entrepreneur: An Economic Theory*, Totowa: Barnes & Noble, 1982.

Caves, Richard, "Economic Analysis and the Quest for Competitive Advantage," *American Economic Review Proceedings*, May 1984, *74*, 127–32.

Gordon, Robert A. and Howell, James E., *Higher Education for Business*, New York: Columbia University Press, 1959.

Griffin, James and Teece, David, *OPEC Behavior and World Oil Prices*, London: Allen & Unwin, 1982.

Hayes, Robert and Abernathy, William J., "Managing Our Way to Economic Decline," *Harvard Business Review*, July-August 1980, *58*, 67–77.

Kreps, D. and Wilson, R., "Reputation and Imperfect Information," *Journal of Economic Theory*, August 1982, *27*, 253–79.

Leontief, Wassily, Theoretical Assumptions and Nonobserved Facts," *American Economic Review Proceedings*, May 1971, *61*, 1–7.

Milgrom, Paul and Roberts, John, "Predation, Reputation, and Entry Deterrence," *Journal of Economic Theory*, August 1982, *27*, 280–312.

Mitroff, I. I., "On Systematic Problem Solving and the Error of the Third Kind," *Behavioral Science*, November 1974, *19*, 383–93.

______, "Towards a Theory of Systematic Problem Solving: Prospects and Paradoxes," *International Journal of General Systems*, No. 1, 1977, *4*, 47–59.

Nelson, R. R. and Winter, S. G., *An Evolutionary Theory of Economic Change*, Cambridge: Harvard University Press, 1982.

Peters, Thomas and Waterman, Robert, *In Search of Excellence*, New York: Harper & Row,

1982.

Porter, Michael, *Competitive Strategy*, New York: Free Press, 1980.

Scherer, F. M., "Technological Change and the Modern Corporation," unpublished manuscript, Swarthmore College, September 1982.

Simon, H. A., *Models of Man*, New York: Wiley & Sons, 1957.

Teece, David J., "Economies of Scope and the Scope of the Enterprise," *Journal of Economic Behavior and Organization*, No. 3, 1980, *1*, 223–47.

______, "The Market for Knowhow and the Efficient International Transfer of Technology," *Annals of the American Academy of the Political and Social Sciences*, November 1981, *458*, 81–96.

______, "Towards an Economic Theory of the Multiproduct Firm," *Journal of Economic Behavior & Organization* No. 1, 1982, *3*, 39–63.

Williamson, Oliver E, "Transactions Cost Economics: The Governance of Contractual Relations," *Journal of Law and Economics*, October, *22*, 1979, 233–61.

______, "The Modern Corporation: Origins, Evolution, Attributes," *Journal of Economic Literature*, December 1981, *19*, 1537–68.

The Values of Economic Theory in Management Education

By ROBERT G. HARRIS*

In their seminal work describing the decline of American industry, Robert Hayes and William Abernathy (1980) identified competitive failures in world markets (loss of market shares at home and abroad); declining productivity (in both absolute terms and relative to Japan and West Germany from 1960 to 1978); and the loss of leadership in both mature and high technology industries. While other commentators had noted the relative decline in American economic performance and cited a large number of alleged causes for this decline, the Hayes and Abernathy article was notable for citing managerial failure as being at the root of the problem.

Although Hayes and Abernathy acknowledged the influence of excessive government regulation and taxation, pressures from labor unions and public interest groups, dependency on OPEC-priced oil, and capital market emphasis on short-run financial returns, they argued that Japanese and West German companies were subject to the same constraints, only more so. How then, they asked, can one explain the poorer performance of American industry by these factors? Instead, they pointed to the "new management orthodoxy" as deserving a major share of the blame, and provided the results of a comparative study of management attitudes in the United States, Japan, and Western Europe to substantiate their charges.[1]

I. The Role of Economics in Management Education

The premise of this paper is that Hayes and Abernathy are basically correct in their assessment of managerial failures. The purpose of the paper is to address a related question: the extent to which business schools have contributed to that failure, with particular emphasis on the role of economics. That business schools have the potential to influence managerial attitudes seems certain: the number of students in undergraduate business programs exceeds 230,000 (Robert Rehder, 1982), and the number of MBA's graduated has increased dramatically from 4,600 in 1960 to more than 62,000 in 1983 (Susan Fraker, 1983). Proof of causality is impossible, but there are very strong correlations between (a) the nature of managerial failures cited by Hayes-Abernathy and Peters-Waterman, and the curricular content of most business schools: and (b) Hayes and Abernathy's period of decline and the period of ascendancy of professional schools of management.

Moreover, the importance of economics and related analytical disciplines have grown markedly in many business schools during the same period. Hence, economics and its practitioners have the opportunity to influence management in three ways:

1) By adopting admissions standards that reflect the growing analytical content of our curricula, business schools play a screening role that influences who will be given the chance to earn a business degree. Since an MBA degree has become a virtual necessity for those aspiring to the managerial fast track (Michael Thomas, 1983), the effects of this screening process are increasing.

2) Once admitted, economists influence students by our attitudes and unstated assumptions in the classroom: the nature of our research (especially when the resulting articles and books are used in courses): and, by example, our behavior in professional activities (William Miller, 1983). Their business school training undoubtedly influences the behavior and attitudes of our graduates as managers and consultants to management.

*Associate Professor, School of Business Administration, University of California, Berkeley, CA 94720.

[1] Thomas Peters and Robert Waterman (1982) also contend that managers and business schools should be assigned a share of the blame for the decline in the competitiveness of American industry.

3) Economists also influence managers, degreed or not, by our published research and our consulting activities. Our business school employers provide an hospitable environment for conducting research; the new management orthodoxy offers opportunities and economic incentives for applying these research methods and techniques in the business world.

In many instances, these effects are quite positive for the individuals and for the performance of the organizations they will or do manage. It is also true that, whatever deficiencies one might cite in management education, they cannot be blamed solely on economists. These caveats notwithstanding, this paper contends that the underlying value assumptions of mainstream economics are at least partially responsible for those deficiencies and their effects on the quality of American management.

II. The Values of Economic Theory

I will review eight of the critical values embodied in economic theory (i.e., the variety of economics taught in most, if not all, business schools). The first of these might be termed a "meta-value," since it concerns the role of values in economics. The next four values relate to the individual: the relationship between personal values and professional practices; the valuation attached to different incentives and objectives; the belief in rational, self-interested behavior; and the preference for analytical over noncognitive skills, and for study over experience. The last three values deal with the corporation and its role relative to other social institutions: the criteria for evaluating managerial and corporate performance; labor-management relations; and an ideological predilection for market "solutions."

It is critical to my argument to distinguish between the values of *economists* and the values of *economics* as a discipline. So far as I know, there has been no systematic study of the ethical or ideological values of academic economists; surely there is considerable variance among us in that respect. But if our professional publications are an indication, there is an extraordinary consensus in what we represent our values to be, whether we "believe" them or not. We teach students these values as we teach economics, especially when—as is often the case—we do not clearly separate the implied normative content of our discipline from what we actually believe.

1) The primary reason why we economists seldom distinguish our own values from those of the discipline is that most of us have been taught that economics is a science, and therefore is, as it should be, value free. While few of us express that view in our classrooms, students learn the principle nonetheless, probably because we devote so little time to fundamental normative issues. As an MBA student put it, "One of the things the Harvard Business School teaches is that there are no truths" (Thomas Moore, 1982, p. 76).[2] Truths there may not be, but the claim that economics is value free is subject to a single interpretation: confusing acceptance of and adherence to a dominant ideology or paradigm with the absence of values.

2) Most economists recognize that individuals do have personal values, but that recognition is conditioned by the assumption that (a) teachers should not impose their own values on their students; and (b) managers need not allow their personal values to constrain their professional behavior. In other words, the conduct of both professors and professionals should be value free.[3] Some educators go so far as to deny any responsibility for dealing with moral or ethical concerns: "As far as ethics are concerned, we figure that our students either have them or they don't" (Moore, p. 76).

The fallacy in the proposition that educators should not impose values lies in the assumption that discussing values, encouraging students to develop and articulate their own values, or considering the implications of personal values for professional conduct is tantamount to "imposing" values. The op-

[2] Although there are a number of references to a recent episode at the Harvard Business School (Moore), the same kind of incident might happen at virtually any business school.

[3] See Peters and Waterman on the importance of, but relative inattention to, values among corporate managers.

posite is true: pretending to be value free, in the face of a value-laden discipline, deprecates the importance and meaning of values to individuals and society.[4]

3) Economics places great value on monetary and material incentives, at the expense of other human aspirations and motivations. Consequently, the discipline fosters the notion that income and wealth are primary measures of the individuals' social worth. This in turn spurs a preoccupation among students with career objectives and advancement.[5] The "career mentality" also contributes to a vocational conception of education, in stark contradiction to the liberal ideal (knowledge for its own sake). The question too many students ask is: what good will it do me? And too many business educators encourage that orientation by concentrating their pedagogical efforts on specialized vocational skills rather than attempting to develop a broader intellectual perspective.

4) A central tenet of neoclassical ideology is the belief that the invisible hand of the market will ensure that rational, self-interested behavior by individuals will optimize social welfare. Yet the current version of that behavioral norm bears little resemblance to its original form: what Adam Smith assumed was *enlightened* self-interest by a community of individuals with common interests and moral principles.

The self-interest norm is also directly at odds with the historical conception and justification of the professions, which recognizes the agent-principal problem: managers are humans, humans are self-interested, therefore managers will attempt to maximize their own interests. The very term professional manager indicates a willingness to exercise self-control and adherence to a set of moral principles, even if contrary to the self-interests of the individual.[6] It is the responsibility of *professional* schools of management to teach that principle.

5) Economics as a discipline values analytical technique and mathematical elegance over informal or institutional knowledge. In fact, economics conceives of management as little more than decisions based on formal theories and analysis. This academic view has found widespread acceptance by pseudo-professionals, who apparently believe that analytical competence is necessary and sufficient for success as a manager; one need have no special knowledge of the industry, products, technology, suppliers or customers (Hayes and Abernathy).

A corollary proposition is that leadership and entrepreneurship are relatively unimportant, or, alternatively, "leaders are born, not made," even though there are discernible differences between managers and leaders, and evidence that, to a degree, leadership skills can be taught (Abraham Zaleznik, 1977). One important consequence of the devaluation of leadership is that too little weight is given to demonstrated leadership ability in the admissions or grading process. Hence, many potential leaders are screened from business programs, namely those applicants who are not analytically inclined or who do not test well.

6) One of the strongest values in economics is that corporate performance should be evaluated according to quantifiable financial criteria, specifically, profits or shareholder wealth. While economists differ as to whether managers optimize or satisfice, there is a virtual consensus that profit is the objective. Although one need not accept profit-

[4] Harvard students contend that "ethical considerations involved in cases studied at the B-school are not often raised by professors.... To ensure that ethical problems are brought up more often in the future, the students decided to add a question to the annual survey grading the business school professors and their courses, asking about the emphasis put on ethics" (Moore, p. 76).

[5] On the basis of a longitudinal study of 220 recent MBA's, Meryl Louis concluded that "even people who work before earning an MBA get caught up in the 'how-am-I-doing' frenzy. MBA programs undermine what people learned on the job. There's this massive resocialization about how much careers matter. People are too bloody busy worrying about two promotions down the road..." (Fraker, p. 72).

[6] In the historical development of the professions, "fears of professional abuse had to be overcome. For this, trust in probity and ethicality of the professional practitioners had to be convincingly established—ethicality being, in Friedson's words, 'prerequisite for being trusted to control the terms of work without taking advantage of such control' " (Magali Larson, 1977, p. 57).

oriented behavior as descriptively valid, it is generally true that managers accept profit as the normative standard for (or rationalizations of) their decisions and actions (Edward Herman, 1981).

That is not to say that profit shouldn't matter, but that too many managers suffer from a myopic view of financial performance, in contrast to the longer-term view that profitability is best attained by investing in technological innovation, new products, better production processes and development of new markets (Hayes and Abernathy). As expressed by Charles Day and Perry Pascarella:

> To improve productivity, then, American management must recommit itself to improving the product and all that stands for...it may be time for such action because the dogma that corporations exist solely for the purpose of generating profits for shareholders is becoming a bit worn... No one is suggesting that business get along without profits. But business also exists to move technology into products and services, and, in turn, move those products and services into the marketplace to improve the society in which it operates and thrives. [1980, p. 55]

Though the longer-run version of profit orientation is superior to short-run myopia, even that perspective denies the legitimacy of other stakeholders' claims on corporate resources and rights of participation in corporate decisions. Yet Edward Freeman (1984) has persuasively argued that denying or ignoring those legitimate interests—employees, suppliers, customers, local communities—can have catastrophic effects on the welfare and survival of the enterprise. Moreover, progressive business leaders recognize the social desireability—even necessity—of making tradeoffs between profit and socially responsible actions (Francis Steckmest, 1982).

7) Another value assumption of economics is the disutility of labor: since work is unpleasant, people must receive adequate material compensation for it. And, since employees work only out of economic necessity, managers must directly supervise their workforce to reduce "shirking." Economics largely ignores other human motivations for working: dignity, affiliation, participation in organized activities, meaning and purpose. But the failure of economics to take account for these diverse motives promotes bad management: "...management's most immediate need is the formation of a new and positive relationship with its workforce, one that accepts employees as partners... and afford[s] employees the chance to find dignity, creativity and relevance in their jobs" (Day and Pascarella, p. 57).

8) Neoclassical economics is essentially the study of markets and behavior in markets. Accordingly, economics exhibits a strong bias for market solutions, and many business school economists exhibit a strong opposition to government intervention in markets. Business students are taught an hostility toward government regulation and taxation in general, and many courses in business and public policy emphasize how managers can exploit their political power and prowess to influence government decisions to their own advantage. While public policies may be misguided or counterproductive, we should acknowledge to our students that, in a political democracy, governments have a legitimate role, even if that imposes costs on business or reduces managerial discretion. A more balanced view of business-government relations might have the desireable effect of reducing the hostility between business and government.

III. Proposals for Reform

While I have been frank in my criticisms, I do not mean to suggest that all is wrong in business education, or that economics is the root of all existing problems. The general quality of business school faculty and curricula has certainly improved over the past twenty years, and many of those improvements were directly responsive to proposals for reforms (for example, see Robert A. Gordon and James Howell's report, 1959) and the needs of the employers of our graduates.

But business education has become a big business, with hundreds of thousands of students, huge budgets and, in some places, cross subsidies to less market-oriented departments. As the invested human capital of our faculties has grown, so too has their vested interest in maintaining the status quo. Further, business schools are as market driven as most corporations, so they can be expected to match their product to short-run customer satisfaction (i.e., the dictates of corporate recruiters), rather than risking possible failure by developing a new, even if greatly improved, product. Still, I believe that there are ample grounds and opportunities for improving the overall quality of, and increasing the value of economics in, management education. We might do this by:

Being more willing to admit our own personal and ideological values to our students.

Identifying the implicit normative assumptions and values in economic theory and discussing their consequences for managers.

Encouraging students to explore, develop, and articulate their own values and discuss their implications for managers.

Decreasing the relative emphasis on analytical aptitudes and training; increasing the value of leadership and entrepreneurship in both admissions and curriculum design.

Teaching respect for experience, while maintaining the value of economic analysis (but as a means, not as an end, of management).

Encouraging students toward intellectual breadth and tolerance, by emphasizing the liberal arts in prebusiness education and liberal attitudes towards professional education.

Increasing resources for development of noncognitive skills, including negotiation, interpersonal communication, and ethical decision making (Rehder).

Educating managers to understand changing values, philosophies, and lifestyles (Rehder), and to appreciate the roles of other institutions in our society (citation to Harvard President Bok's Report on Harvard Business School, Moore, p. 76).

REFERENCES

Day, Charles R. Jr. and Pascarella, Perry, "Righting the Productivity Balance," *Industry Week*, September 29, 1980, 50–59.

Fraker, Susan, "Tough Times for MBA's," *Fortune*, December 12, 1983, 65–72.

Freeman, R. Edward, *Strategic Management: A Stakeholder Approach*, Boston: Pitman, 1984.

Gordon, Robert A. and Howell, James E., *Higher Education for Business*, New York: Columbia University Press, 1959.

Hayes, Robert H. and Abernathy, William J., "Managing Our Way to Economic Decline," *Harvard Business Review*, July-August 1980, *58*, 67–77.

Herman, Edward S., *Corporate Control, Corporate Power*, Cambridge: Cambridge University Press, 1981.

Larson, Magali S., *The Rise of Professionalism: A Sociological Analysis*, Berkeley: University of California Press, 1977.

Miller, William E., "The Untaught Skill," *Collegiate News and Views*, Winter 1983, *36*, 19–23.

Moore, Thomas, "Industrial Espionage at the Harvard B-School," *Fortune*, September 6, 1982, 70–76.

Peters, Thomas J. and Waterman, Robert H. Jr., *In Search of Excellence*, New York: Harper & Row, 1982.

Rehder, Robert R., "American Business Education: Is It Too Late to Change?," *Sloan Management Review*, Winter 1982, *23*, 63–71.

Smith, Adam, *The Wealth of Nations*, London: Methuen, 1904.

Steckmest, Francis W. (with a review and resource committee for the Business Roundtable), *Corporate Performance*, New York: McGraw-Hill, 1982.

Thomas, Michael M., "Business Education: A Study in Paradox," *Business and Society*, Spring 1983, *22*, 18–21.

Zaleznik, Abraham, "Managers and Leaders: Are They Different?," *Harvard Business Review*, May-June 1977, *55*, 67–78.

Economic Analysis and the Quest for Competitive Advantage

By RICHARD E. CAVES*

An extensive meeting of interests has occurred in the past decade between applied microeconomists and those who study business strategy. Business strategists have cast up a series of questions for industrial economists, who have responded with numerous applications of economic theory and quantitative research methods. This interchange has been fruitful, and indeed more fruit remains for the picking. I shall argue for centering the collaborative effort on what may be called "committed competition," rivalrous moves among incumbent producers that involve resource commitments that are irrevocable for nontrivial periods of time. Most concerns of business strategy are with policy choices that involve some degree of commitment. The economic analysis of market processes, conversely, has until recently confined itself to competitive moves assumed to be uncommitted and reversible. The assumption is reasonable enough when we consider short-run price-output determination in a market subject to precommitted capacity constraints. It is not so reasonable when the irrevocable decisions are being made under rivalry.

I shall argue that most significant business decisions involve substantial commitments, with attendant uncertainties, information problems, and first-mover advantages, and that business strategists have correctly sought to illuminate the quest for competitive advantage. Economic analysis until recently has avoided this imperative (for reasons inherent in our standard comparative-statics methodology). The first section considers the relevance of committed competition from the vantage point of business strategy, while the second investigates what economic analysis has to say about it and where further contributions may lie.

I. Committed Competition and Business Strategy

Prominent writings on business strategy for practitioners, such as Michael Porter (1980, ch. 2), identify certain strategic options for a firm seeking advantage over its competitors. Porter's list allows the firm basically two choices: it can seek lower costs than its competitors; or it can differentiate its product and auxilliary services in a number of ways that entail providing customers with a higher quality option than do competitors. Cutting across these options is a "focus" strategy that caters to a particular niche afforded by buyers' heterogeneous preferences. Unless the firm has the niche to itself, it still faces the same options for gaining advantage against rival occupants of that niche.

The important question is not whether this array of strategic options is substantively complete, but why the options are viewed as strategic and exclusive. Each involves a number of specific investment decisions—construction of efficient-scale facilities and vigorous pursuit of cost reduction through accumulated experience, in the case of cost leadership; research and design, quality control, customer-service facilities and the like, when differentiation is pursued. These discrete investments are generally not mutually exclusive. However, a sufficient source of exclusivity lies in managerial coordinating capacity and the need to select a system of internal organization, evaluation, and reward that is designed for optimal pursuit of the chosen strategy. That is, a firm's managerial cadre may hope to beat its median-ability competitor along one dimension, but not along every dimension.

This view of competitive options draws support from the prevailing business-normative definition of corporate strategy. Put in economic terms, the definition presumes that the firm operates in the short run and is contractually encumbered with a variety of

*Littauer Center 210, Harvard University, Cambridge, MA 02138. Suggestions by Therese Flaherty and Pankaj Ghemawat are gratefully acknowledged.

fixed facilities, including a staff of specialized personnel and assorted types of firm-specific knowledge. Strategic choice then expresses the top coordinator's attempt to maximize the rents to these fixed factors over the planning horizon (the objective function may hold other arguments besides profit). Strategy gives rise to organizational choices and to decisions about acquiring and divesting assets that are selected for the efficient pursuit of this maximization plan. This normative formulation is backstopped by an impressive amount of behavioral evidence ranging from the historical studies of Alfred D. Chandler, Jr. to a variety of statistical studies affirming the positive value of choosing a strategy that is correctly matched to the attributes of the firm's market and making organizational choices that best serve the elected strategy.[1]

This view of competitive options may seem to clash with economic analysis on the degree to which business decisions are irreversible over appreciable periods of time. That question is simply empirical. To appreciate the element of truth in the business-strategy view, we can note the decision element of an investment-type commitment behind practically any sort of competitive move, with resources sunk in place before the state of nature and the rivals' reactions are known. Some of these commitments are long-lived and irreversible—specific to the firm and limited in resale value. This is obvious in the case of capacity expansion that must precede a price cut,[2] or the research and development that supports introduction of an improved product. Changes in product quality or design require prior alteration of production facilities and a buildup of inventories of the revised product. Changes in ancillary services demand the recruitment and organization of the necessary personnel and/or contracts with independent distribution or service firms. Other competitive moves, while yielding no deposit of tangible capital, nonetheless require extensive fixed outlays preceding the actual move; advertising campaigns or other informational outlays must implement any general change in price or nonprice dimensions of the firm's offer. Even pricing strategies have an investment aspect, because they normally encounter some response lag on the part of buyers as well as any fixed costs of announcement and capacity expansion. In short, one can argue that business competitive moves generally have intertemporal investment aspects. The implied depreciation rates of the investment components range from short to very long. Salvage values, in the case of abandoned strategies, may be positive in some cases but zero for important classes (research, sales-promotion outlays).

Thus, the business-strategy approach correctly presumes that many competitive moves are strategic in the sense of involving substantial precommitments of resources, and that they are exclusive in the sense that a firm is unlikely to possess the organizational capacity to beat its median competitor in more than one strategic dimension.

II. Applications of Economic Analysis

How to characterize strategic rivalry as an economic process seems straightforward enough, and indeed we know a lot about it as single play games in which one allocation is to be made subject to strategic preemption. What we may lack is analysis, both theoretical and empirical, of markets in which a series of investment commitment opportunities present themselves over time to a set of incumbents, such that the initial conditions for any given round reflect the undepreciated residue of commitments made in past periods. This section considers what we know and how the pieces relate to business strategy.

Economists' theoretical interest in commitment has been largely in the context of entry barriers and entry deterrence. Despite the hubbub over the concept of contestable markets, any charitable reading of the theory of entry barriers from Joe Bain (1956) on makes it clear that each source of entry

[1] More detail appears in my 1980 survey of this literature.

[2] Richard Smith (1981) showed that investment decisions are sensitive to a wide range of competitive considerations. These considerations pick up the sensitivity of the firm's expected profit to its rivals' actions and the uncertainty of those actions.

barriers rests on some irrevocable resource commitment by incumbents in the market. Entry barriers are therefore barriers to exit as well, and any rent-seeking behavior undertaken by incumbents to shrivel the expected profit opportunities for entrants has the side effect of increasing the sunk resources that the incumbent has at risk in some adverse state of nature. Research on entry deterrence has proceeded to identify a number of dimensions in which it might pay to expand resource commitments beyond their myopic optima in order to make entry infeasible or less likely. These dimensions of commitment include production capacity, longevity of capital equipment, sales-promotion outlays, product strategies such as brand proliferation and model changes, vertical integration, contractual commitments (with penalties) to customers or suppliers, research outlays of certain types that preempt intangibles, and cost-raising strategies that differentially disfavor entrants.

A good deal of research has established conditions under which entry may be feasibly deterred by expanding these outlays. Interest has now turned to the complementary question of whether deterrence is profitable —that is, whether the profits of the incumbent(s) net of deterrence costs exceed those that would be expected in a post-entry situation—specifically, in a post-entry equilibrium that the incumbent expects would be reached.

While the research on entry deterrence holds considerable interest, it undeniably addresses quite special circumstances. Because entry barriers are a private collective asset of an industry's incumbents, investments to augment them are subject to free riding and underprovision. While structural barriers to entry unquestionably keep down the inflow of entrants (given the inducements to commit additional resources), we have no convincing evidence that the resulting loss of welfare triangles is greatly amplified by significant rent-seeking investments by incumbents. However, if incumbents may seldom possess the information and cohesion needed for tuning investments to deter entrants, there are certainly enough markets where recognized interdependence may warrant strategic actions aimed at extant rivals. Incumbents' reaction functions may be known, whereas potential entrants' proclivities surely are quite uncertain. And with spatial elements in markets frequently making some incumbents closer rivals than others (more on this below), behavior to deter rival incumbents should be a more active option than conscious commitments to deter entrants.

Can these strategies for analyzing entry deterrence be turned to the analysis of rivalry with recurring commitment opportunities? The answer seems to be affirmative, and some progress has been made. One contribution feeding directly on the literature of business strategy is the concepts of strategic groups and mobility barriers. The strategic-group concept assumes that firms make committed strategic choices, and that heterogeneities of technology and demand even in well-defined markets may make several different strategies viable (i.e., yield at least normal profits conditional upon other strategic niches having been occupied). Firms following similar strategies are likely to be sensitized to each others' actions by high cross elasticities, making them more likely to anticipate reactions by their group rivals than by other incumbents. Furthermore, the factors delineating strategic groups themselves are directly related to structural barriers to entry, establishing a straightforward economic explanation why some strategies prove persistently more profitable than others in the same market. Entry barriers thus proliferate as barriers to mobility among strategic groups, and deterrence investments can take the form of actions to reduce the profit opportunities of rivals who would match the firm's strategy.

Empirical research has supported this framework for analyzing committed competition. Howard Newman (1978) found that strategic groups could be delineated on the basis of incumbents' similarities in diversification and vertical integration, and that a more complex group structure in an industry makes its aggregate profits both lower and less predictable than otherwise.[3] Porter (1979)

[3] Michael Hergert (1983) demonstrated more broadly the empirical bases for strategic groups, their determinants, and their net effect on market performance.

assumed that similar-size firms in an industry are more likely to share common group affiliations; he confirmed the resulting prediction that the profits of firms in different size-classes should be sensitive to different structural determinants. Sharon Oster (1982) showed that strategic differences based on advertising rates are persistent, and persist longer in industries where the depreciation rates of advertising are lower. Martin Ramsler (1982) similarly affirmed the persistence of strategic differences and showed that they predict which firms will seize subsequent investment opportunities as they occur. My study with Thomas Pugel (1980) reported that simple descriptors of strategic differences within industries can successfully predict differences in overall market structure and in the interfirm (intra-industry) distribution of profits. Some of these statistical studies have been replicated abroad, and case studies offer additional confirmation.

The concepts of strategic groups and mobility barriers do not add up to a tight formal model. Rather, they serve to organize predictions that come from tight models and assist confronting them with empirical evidence—a dynamized add-on to the traditional structure-conduct-performance paradigm. What of the formal models themselves? We are beginning to explore the processes of sequential commitment. The explorations have both business normative (how can the manager maximize his objectives?) and welfare-normative sides. The following paragraphs suggest the sorts of business-normative conclusions that are emerging.

Some contributions directly highlight the managerial decision in committed competition. Pankaj Ghemawat (1983) has adapted the "'winner's curse" of auction theory to industrial investment with incomplete information. He showed that, when the market offers room for a single new plant, firms may bid implicitly for the earliest time at which they will construct it; if the winner's curse operates, the victor is the firm that most overestimates the rate of growth of demand.[4] Another way to formulate this model is in terms of the penalty of unalterably suboptimal scale that a firm will endure in order to preempt an investment opportunity—a form of rivalry that seems to dominate in the evidence of one empirical study (Ronald Johnson and Allen Parkman, 1983). The penalty of excess cost for the plant's lifetime is an alternative to the penalty of carrying excess capacity temporarily.

Other problems of incomplete information are highlighted by models that assume a firm must independently make some of its outlays on committed investments before it knows whether or not it has beaten out its rivals. If these outlays are not preemptive and information is incomplete, then no mechanism exists to align the sum of precommitment outlays to a privately or socially optimal level. Economic analysis has recently turned to models of a race to attain some preemptive commitment—to acquire a piece of information, make an innovation, build a plant, establish a predominant standard or configuration, etc. With full information and certainty, either somebody precommits, preempting outlays by rivals, or a standoff occurs and nobody starts a race that only one can win. With uncertainty, "leap-frogging" opportunities, and the like, the outcomes of commitment races (and appropriate strategies for those entering them) depend on extensive information (including information about one's potential rivals) and subtle evaluations of probabilities (Drew Fudenberg et al., 1983).

Incorporating the implications of these uncertain commitments into business decision rules may have a substantial payout in improved decision making. Amos Tversky and Daniel Kahneman (1978) supplied impressive evidence that even sophisticated decision makers underutilize information that is available to them on prior probabilities and rely excessively on evidence from the instance at hand. For example, the project at hand that "looks good" puts out of consideration the small proportion of such projects that have succeeded in the past. Their findings certainly underline the relevance of the winner's curse.

Potentially the richest source of business-normative implications for committed com-

[4] This model offers the testable proposition that industries in which investment opportunities are lumpy and discrete will, other things equal, earn lower profits than those whose commitments come in smaller lumps (Sidney Schoeffler et al., 1974, offer suggestive evidence).

petition is the fast-proliferating closed-loop models, which allow commitments to be revised (but not totally unwound) in light of market results. Some of these models deal with entry deterrence, others with committed rivalry among incumbents. The latter group, of interest here, are technically complicated —combining as they do sequential decisions under uncertainty with equilibrium conditions in oligopolistic markets. These conclusions, stated in positive and normative economic terms, have their strong business-normative implications as well. I sample some positive predictions about market structures, then consider their business-normative implications for rivalry among incumbents.[5]

Some of these models predict a shrinking variance of firm sizes, because the smallest or most disadvantaged firm has the most to gain from preempting the commitment opportunity. Richard Gilbert and Richard Harris (1981) found that the smallest firm is most likely to preempt an investment, because it has the least to lose from any resulting price reduction on its previous output. Jennifer Reinganum (1983) got a somewhat similar result for the adoption of an innovation that is not necessarily profitable if all firms adopt it. On the other hand, Richard Nelson and Sidney Winter (1978) indicated from a similar model based on simulation methods that concentration tends to increase with an industry's latent rate of productivity growth (which expands the variance of the outcome of successful innovation) and the difficulty of imitating innovations. They also observed that initially high seller concentration tends to retard the growth of concentration through this random process, because it limits the innovator's opportunities for profitably increasing his market share. Clearly, the structural consequences of committed competition can be diverse, depending on the size and longevity of the returns to a successful preemption and what differently situated firms may expect to gain from a preemptive move.

These models offer evident business-normative implications—how the firm can assess its relative advantage in pursuing some opportunity as well as what degree of competition will prevail after successive rounds of commitment in its market. However, one may wonder whether the exclusionary commitments targeted by the closed-loop models are really the same thing as the competitive advantages sought through strategic focus. After all, the business-normative literature stresses the pursuit of rents that rival firms are incapable of attaining, and not the pursuit of rents by pre-empting their attainment. The difference, I submit, matters for welfare economics but not for business behavior. The rent (i.e., advantage) commanded by a strategy declines with increases in the number of rivals that can replicate it and their reaction speeds. A strategy can be profitable either because it amounts to a natural monopoly (blockaded entry) or because its adroit implementation by a limited number of firms makes its replication unprofitable for latecomers. We are back to the issue of entry deterrence: the incremental profitability of commitments that widen a strategic advantage is greater, the more they reduce the chances of replication and the more intensive is the rivalry that will result if the strategy is in fact replicated.

The one element of difference flagged by the business-strategy view lies in heterogeneous assets with special qualities that are contractually attached to the firm, *by assumption* not available to would-be preemptors. That successful firms possess such assets is assumed by some researchers to explain the observed correlation between firms' profits and market shares. Successful strategies based on such assets may avoid impairing economic welfare through strategic preemption, but they do raise instead the question of why the superior productive asset neither appropriates the rent due to its differential ability nor attracts bids to join other coalitions on terms that would award it these rents.

REFERENCES

Bain, Joe S., *Barriers to New Competition*, Cambridge: Harvard University Press,1956.

Caves, Richard E., "Industrial Organization,

[5] David Kreps and Michael Spence (1983) provided a more complete analysis of this literature as well as related aspects of intertemporal competition.

Corporate Strategy and Structure," *Journal of Economic Literature*, March 1980, *18*, 64–92.

______ and **Pugel, Thomas A.**, *Intraindustry Differences in Conduct and Performance*, No. 1980-2, GSBA, New York University, 1980.

Fudenberg, Drew et al., "Preemption, Leapfrogging and Competition in Patent Races," *European Economic Review*, July 1983, *22*, 3–31.

Ghemawat, Pankaj, "On Competition, Chance, and Capacity Expansion," manuscript, Harvard Business School, 1983.

Gilbert, Richard J., and Harris, Richard G., "Investment Decisions with Economies of Scale and Learning," *American Economic Review Proceedings*, May 1981, *71*, 172–77.

Hergert, Michael L., "The Incidence and Implications of Strategic Grouping in U.S. Manufacturing Industries," unpublished doctoral dissertation, Harvard University, 1983.

Johnson, Ronald N. and Parkman, Allen, "Spatial Monopoly, Non-Zero Profits and Entry Deterrence: The Case of Cement," *Review of Economics and Statistics*, August 1983, *65*, 431–39.

Kreps, David M. and Spence, A. Michael, "Modeling the Role of History in Industrial Organization and Competition," Discussion Paper No. 992, Harvard Institute of Economic Research, 1983.

Nelson, Richard R., and Winter, Sidney G., "Forces Generating and Limiting Concentration under Schumpeterian Competition," *Bell Journal of Economics*, Autumn 1978, *9*, 524–48.

Newman, Howard H., "Strategic Groups and the Structure-Performance Relationship," *Review of Economics and Statistics*, August 1978, *60*, 417–27.

Oster, Sharon, "Intraindustry Structure and the Ease of Strategic Change," *Review of Economics and Statistics*, August 1982, *64*, 376–83.

Porter, Michael E., "The Structure within Industries and Companies' Performance," *Review of Economics and Statistics*, May 1979, *61*, 214–27.

______, *Competitive Strategy*, New York: Free Press, 1980.

Ramsler, Martin, "Strategic Groups and Foreign Market Entry in Global Banking Competition," unpublished doctoral dissertation, Harvard University, 1982.

Reinganum, Jennifer F., "Technological Adoption under Imperfect Information," *Bell Journal of Economics*, Spring 1983, *14*, 57–69.

Schoeffler, Sidney et al., "Impact of Strategic Planning on Profit Performance," *Harvard Business Review*, March/April 1974, *52*, 137–45.

Smith, Richard L. II, "Efficiency Gains from Strategic Investment," *Journal of Industrial Economics*, September 1981, *30*, 1–23.

Tversky, Amos and Kahneman, Daniel, "Judgment under Uncertainty: Heuristics and Biases," in P. Diamond and M. Rothschild, eds., *Uncertainty in Economics*, New York: Academic Press, 1978, 19–34.

Reform of the Budget Process

By ALICE M. RIVLIN*

Budget decision making at the federal level in the United States can hardly be described as casual, haphazard, or ill-informed. The budgeting process is lengthy and elaborate. The principal decision makers—the president and the Congress—are assisted at every stage by an army of highly trained economists and budget analysts who assemble masses of information, use sophisticated forecasting models, and have access to state-of-the-art computers. Everyone works very hard. No government in the world devotes as much time, energy, and talent to budget decision making as our's does.

Nevertheless, almost everyone is unhappy, with both the outcome of all of this effort and the process itself. Of course, some dissatisfaction with the outcome is normal. No matter how smoothly the mechanics of budget decision making are carried out, some will feel that the government spends too much or too little, spends on the wrong things, or taxes in the wrong way. But the current situation is not normal. Unless recent budget decisions are changed, they will lead to high and rising structural deficits that almost no one defends as desirable fiscal policy, a rapidly escalating burden of debt service, and a mix of fiscal and monetary policies that is reducing U.S. competitiveness in international markets and seems likely to retard growth.

Moreover, quite apart from its unsatisfactory outcome, both participants and observers decry the shortcomings of the budget-making process itself. Budget decision documents are complex, technical, and difficult to understand. Even the experts have a hard time following what is going on. There never seems to be enough time for debate or deliberate decision making at any stage of the process, but the whole process takes too much time. Executive officials and members of Congress seem to do nothing but defend, question, and debate the budget and still they never finish before the budget year begins—and sometimes not before it ends. On top of all this, the economic assumptions are always proving wrong. Decisions are out of date almost before they are made. It is all very frustrating.

It is tempting to ask whether there are not some procedural reforms that could solve all of these problems. Couldn't we change the budgeting process so that it would be easier to understand, less time consuming, less uncertain, and, above all, less prone to produce large budget deficits?

I will argue that our current problems are not primarily procedural. The budgeting process is complex and time consuming primarily because the federal government does so many different kinds of things, and because Congress is so reluctant to concentrate on major directions of policy while leaving the details to executive departments or state and local governments. We can simplify the budget process only by simplifying the government itself and changing the role of the Congress. We can make the budget process less time consuming only if we are willing to make decisions less often, or to give up some checks and balances. Moreover, the world is an unpredictable place, and, while we could perhaps handle unpredictability in the budget process better than we do, no procedural changes can eliminate it. Nor does the failure to make the hard decisions neces-

*Director, Economics Studies Program, The Brookings Institution, 1775 Massachusetts Avenue, NW, Washington, D.C. 20036. The views expressed in this paper are my own and should not be ascribed to the officers, trustees, or other staff members of the Brookings Institution.

sary to bring budget deficits down reflect biases built into our budget-making procedures. It is simply a failure of political will and national leadership which an alternative set of procedures will not remedy.

I. The Complexity Problem

At first glance, it is not obvious why making a budget for the U.S. government should be so complicated. On the spending side, why is it not possible to list what was spent for each activity of the government last year, argue about priorities, decide whether to spend less or more depending on perceived need, whether the money is being effectively spent, and whether something else is more important? Why not then set aside a specific sum for each activity and scale back these sums if the total seems to be getting too large? The tax rates could then be set so as to produce the desired amount of revenue.

The realities of making a budget are more complicated than this for at least three reasons. First, much of federal spending goes to individuals for certain specified reasons, such as retirement, ill health, and poverty. Legislative efforts to distribute these benefits fairly and prevent abuse have given us an extremely complicated set of rules governing entitlements. For any given set of rules, the funds expended will depend on the state of the economy, how many beneficiaries apply, and what the rules say they are eligible for. There is no way to change the level of funding without changing some of these rules and estimating the effect of the rule changes on future funding.

It is possible, of course, to rewrite the rules to shift some of the risk of unexpected events from the government to the beneficiary of the programs. For example, if prices rise faster than wages, the lower of the two indexes can be used to determine pension benefits. But these kinds of rule changes do not make budgeting for entitlement programs simpler. There is no way to make it simpler without changing the nature of the relationship between the government and the beneficiary.

Second, for many government activities (especially procurement and construction), money allocated for a project is actually spent over several years at rates that are only roughly predictable. This means that the budget maker is always dealing with two sets of numbers for any particular year: the amounts to be allocated for new projects and the spending that will occur from projects already undertaken. It is difficult to make changes on short notice, especially in defense where a high and currently rising proportion of outlays in any one year are from prior year appropriations.

The third and most important source of complexity in federal budgeting is simply that carrying out the general purposes of the national government involves a myriad of detailed actions. Congress over the years has concentrated on controlling these details through budgetary action rather than confining its role to spelling out major strategies and purposes. In the domestic arena, the result has been a plethora of grants to state and local governments for narrow "categorical" purposes rather than more general intentions like "improving education." Some of these narrow programs have been consolidated in recent years, but thousands remain. They do not account for nearly as many dollars as entitlement programs, but they contribute substantially to the complexity of the budgeting process.

In the national security arena, congressional concern with detail drives out consideration of the major objectives of national security policy and how they are being met. The administration presentation of the defense budget could be an occasion for an examination of our defense posture, the nature of the perceived threat to our security, what U.S. capacity is or could be to deter or repel that threat, and the state of readiness of the various kinds of forces that might be deployed in the event of emergency. Instead, Congress writes very detailed rules about weapons systems and facilities and endeavors to control a lot of details, but fails to demand an explanation of the grand design into which all of this detail is supposed to fit.

The result is that the Budget Appendix is as long and as readable as the Manhattan phone directory. Authorization, appropriation, and tax bills are voluminous and complicated. Nor is the preoccupation with minutia confined to authorizing, appropriation, and tax committees. Debate on the budget resolution, which is supposed to provide an opportunity for weighing the importance of one major set of activity against another, often ends up focusing on detailed concerns about protecting very specific programs.

There will be little progress in reducing the complexity of the budgeting process until Congress realizes that it can contribute more to changing the future of the nation by directing overall policy, than by controlling details. Congress would have to aspire to be an effective board of directors, rather than an ineffective national management.

II. The Time Consumption Problem

If the public perceives that Congress and the Executive Branch spend most of their time arguing about budget issues and never get the decisions made on time, it is only because it's true. In the Reagan years the agonizing budget debate has virtually eclipsed other governmental activity. This is partly because the Reagan program, which Congress passed in 1981, was such a drastic break with the past, and partly because the existence of the looming budget deficits, which this program created, poses extremely unpleasant choices about which there is as yet no national consensus.

But even in more normal times, the budget decision calendar can only be described as crazy. In principle, the president and his budgeteers are supposed to work with all the departments and agencies of government through the fall months on a detailed plan for all the government's spending programs in the next year and for taxes as well. The Congress gets this plan in January and works through the spring, holding hearings, gathering information, and consulting experts, so that by May it can finish up any legislation needed to authorize spending programs and act on an overall budget plan for the coming fiscal year in the form of a First Concurrent Budget Resolution. The summer is supposed to be devoted to bringing detailed spending and taxing legislation into line with that plan, and September to a second and final budget resolution which reconciles individual programs and taxes with that budget. Then the budget year begins on October first, and the whole process of budgeting for the next year starts again.

In practice, this demanding decision schedule cannot be met. Even with the Congress—very sensibly—dispensing with the second budget resolution and starting the reconciliation process earlier, everything runs late. Appropriations bills do not get finished by the time the budget year starts and part of the government runs on continuing resolutions, which certainly looks messy, although it has little real significance.

Possible solutions to the time consumption problem fall into two classes: make decisions less often or dispense with some of the stages. If the government could make its budget decisions every other year instead of every year, the decisions themselves would not be any easier, but the intervening year could be used for something other than budget debate. Moreover, program managers would have more time for managing, and state and local governments would have a longer planning horizon for federally funded programs, and might operate those programs more effectively.

The main trouble with this is that forecasts would have to be made for two years instead of one, and everyone would have to resist the temptation to reopen budget decisions when conditions changed. If biennial budgeting simply meant that every program got a supplemental appropriation in the intervening year, it would accomplish nothing. To release time for other activities, Congress would have to submerge its disposition to tinker with the rules, and trust program managers for two whole years. Managers would have to plan ahead more and reject excuses to run back for more money in less than a dire national emergency.

The number of stages in the budgetary process could be reduced if the Executive Branch and the Congress joined forces to explore how programs were working and consider budgetary alternatives. But this seemingly practical consolidation is fundamentally threatening to the separation of powers. The role of the president in our system is to present and defend his program; the role of the Congress is to debate that program and decide (subject to veto) what to do. To alter the sequence is to alter the role of the Congress and the president in ways that ought not to be undertaken lightly.

A less drastic time saver, which involves only the organization of the Congress, would be to consolidate the authorization and appropriations processes. It is not obvious why we need one committee to prepare a bill to authorize a program and another to recommend how much to spend on it once it is authorized. This duality is a check and balance we may no longer be able to afford.

The point is that making the budgetary process less time consuming involves more than superficial changes in the calendar of budget decision making. It involves fundamental alterations in the way the government works.

III. The Problem of Economic Uncertainty

One of the most frustrating aspects of budget decision making is that the numbers jump around so much. The budget is a plan for the future, and hence, must necessarily be based on a whole set of assumptions about what the future will be like. No one argues much about the noneconomic assumptions. It is implicitly agreed that the weather will be average, and that the level of international tension will not rise enough to matter. The arguments relate to assumptions about economic growth, inflation, and interest rates. Even small differences in the assumptions make big differences in budget magnitudes especially after several years.

Often there are competing assumptions. The administration may put forward a budget proposal which appears to lead to a balanced budget on optimistic assumptions about the future course of the economy. A competing proposal, which actually incorporates higher tax rates, may appear to lead to a substantial deficit because it is based on less optimistic economic assumptions.

Even if initial assumptions are agreed to, the budget decision process is very long, and economic forecasters frequently change their views in the middle, and introduce new and different assumptions. Additional changes are likely to occur as the budget year wears on. The final numbers may look very different from the ones on which Congress voted, even if no policy changes have been made. It is hard for a politician to summon up the courage to vote for a tax increase or a spending cut on the grounds that such action is needed to reduce the deficit when he knows that the projected deficit may be substantially reduced if the economy grows more strongly than predicted, or may be very much larger if the economy gets worse. Under such circumstances, acts of courage get very little reward.

Only part of this problem is fixable. It would certainly be easier for the administration and the Congress to argue intelligently and constructively about the budget if they could agree on a common set of economic assumptions at the beginning of the budget decision process. In most years, it would probably not be hard to reach such an agreement, since, despite our profession's reputation for dissension, the economic forecasting community usually clusters tightly around a consensus view and differences are not consistently ideological. The Reagan Administration's exceedingly optimistic forecast of 1981 was unprecedented and has not been repeated. In case a dispute did arise, a neutral party could be called in to resolve it—perhaps the Federal Reserve. It would be useful to make this common set of economic assumptions moderately pessimistic—not necessarily a worst case scenario, but one from which the deviations seem likely to be in the nature of pleasant surprises.

Agreement on budget assumptions, of course, does not alter the fact that any set of assumptions is likely to prove wrong. This fact just has to be lived with and probably does not even matter much, as long as it is not used by politicians as an excuse for inaction.

IV. Dealing With the Deficits

Before the passage of the Budget Reform Act of 1974, it could be argued that congressional budget procedures had a pro-spending and pro-deficit bias. There were pressures to spend from beneficiaries of programs and pressures to cut taxes, but no moment of truth in which the Congress had to come to grips with the budget as a whole, put the spending and revenue totals together, and go on record as favoring a deficit. But now the Congress spends much of its time agonizing over the budget aggregates and the prospective deficits. Moreover, under the leadership of the Reagan Administration, Congress has made drastic cuts in some of its favorite programs—the narrow grants and local public works that give a congressman his visibility and a subcommittee chairman his power. Spending is no longer growing for the old pork barrel, log-rolling reasons. Prospective spending growth is concentrated in a small number of programs with very broad popular support (primarily defense, pensions, and medical benefits), and revenues after the tax changes in 1981 and 1982 are simply not growing fast enough to close the gap, even if the economy continues to improve. This is not a procedural problem, it is simply a question of wanting more government services than there are revenues to pay for them and having to make some hard choices to bring the two sides of the budget closer together.

Even giving the president an item veto does not seem likely to help much in reducing deficits. The president might veto a few individual spending items, but these are unlikely to affect the totals appreciably. An item veto cannot be used to reduce current entitlement programs, and, while the president might be tempted to close a few military bases, he would be reluctant to jeopardize support for his defense spending increases by alienating pro-defense congressmen. An item veto would help reduce the deficit only if an important part of the deficit were attributable to spending for small items that the Congress wanted and the president didn't. This simply is not the case. When the deficit arises from growth in a few big programs outrunning growth in revenues, there is no solution except to cut the growth in those programs or raise taxes, or both. Procedural changes will not make these choices easier.

Which Budget Deficit? Some Issues of Measurement and Their Implications

By ROBERT EISNER*

A budget deficit is like sin. To most of the public it is morally wrong, very difficult to avoid, not always easy to identify, and susceptible to considerable bias in measurement.

To the body politic, and perhaps also to many economists, the apparent underlying reality is that every dollar of deficit—of a person, business or government—adds a dollar to debt. And debt is bad!

In commentary on that last, I like to recall pleasurably the late Sumner Slichter, a respected conservative economist of at least a generation ago. Churlishly responding to a pressing questioner on a TV program, he suggested increasing debt could be very good; he would generally counsel young people to go into significant debt to buy homes.

More- and less-sophisticated economists have long come to recognize a number of measures of federal budget deficits and their varied implications. Deficits are not always bad, or at least at certain times their alternatives are worse. And we must distinguish among cyclical deficits, structural deficits and high employment deficits, to name a few. But matters are more complicated than many of us have thought to note. We do not usually measure government budget deficits in economically relevant fashion. When we do, we can get results at variance from much conventional wisdom. Just consider the following sketchy set of issues.

1) Should we have included in the fiscal 1983 budget deficit of $195.4 billion some $17 billion of outlays of off-budget federal entities largely to finance direct loan programs? In fact, we did not.

2) As of the end of the 1982 fiscal year, the Treasury listed "contingency" obligations which totalled $6,982 billion, largely for retirement pay and Social Security. Should we try to estimate the year's increase in present value of obligations, net against them the increase in present value of anticipated associated receipts and add that remainder to the deficit? We do not.

3) In general, when the federal government incurs a liability in order to acquire an asset, whether borrowing to make a loan or to acquire real capital, should we net the additional asset from the additional liability in calculating our deficit? Federal investment-type outlays have been estimated by the Office of Management and Budget at $182 billion fiscal 1983. Should the federal government, like private business and state and local governments, have separate capital and current budgets? It does not.

4) We all know that inflation plays huge tricks on conventional accounting. Paul Pieper and I (1984) have estimated that adjusting for changes in the real market value of federal financial assets and liabilities would change the high-employment budget surplus by large *and varying* amounts, ranging between 2.86 percent and −0.18 percent of *GNP* in the years 1969 to 1982. Adjustments would be broadly similar for actual budget deficits. Should we make inflation adjustments? We do not.

5) The $195.4 billion deficit for fiscal 1983 cited above, we are told, is a historical record for the United States. We have never had a budget deficit that large. And it is anticipated by the Congressional Budget Office that if no corrective action is taken (even assuming that defense spending will grow at a rate of "only" 5 percent after inflation, lower than administration requests) the deficit will rise to about $250 billion in fiscal 1988 and $280 billion in 1989. But so what? What do these mind-boggling numbers mean? Gross National Product in calendar 1983 will be over $3,300 billion

*William R. Kenan Professor of Economics, Northwestern University, Evanston, IL 60201. I am indebted to Paul Pieper for his contribution to joint work on which this paper leans heavily.

and, by 1989, according to one set of forecasts, it will be in excess of $5,400 billion. Should we perhaps recognize that in an economy in which everything is growing—output, income, liabilities and assets—a quite substantial deficit may leave ratios of financial assets and liabilities in our government and national balance sheets unchanged? We generally do not.

6) With all the questions relative to knowing what budget deficits are or have been, problems are vastly compounded in forecasting deficits in budgets yet to be realized. As is well known, huge portions of federal receipts and expenditures are endogenous, determined not merely by decisions as to rates of expenditures and taxes, but by economic and demographic developments, and, indeed, interactions between the budget and the economy. We can perhaps forecast how many people will be in each category eligible for Social Security, but we do not know how many will choose to draw benefits rather than continue working. What assumptions should we make about employment, unemployment, real growth, inflation, corporate profits, and so many other lesser variables which will so critically affect future budget deficits? Should we have a single forecast of these variables, agreed upon by Congress and the administration, so that differences in deficit estimates and predictions can be separated from differences in underlying assumptions? We do not.

7) Finally, how do differences among the measures of budget deficits, past and prospective, relate to their possible impact on the economy? Have we considered which are relevant for the various issues of so much concern—inflation, employment, investment, consumption, distribution of income, role of government, incentives, efficiency and the allocation of resources?

Having sketched out a set of questions, I shall be so bold as to offer some comments, including even a few numbers on which to reflect.

I. Off-Budget Items and Credit Extension

Off-budget items relate almost entirely to credit extension, generally by the Federal Financing Bank to federal lending agencies. It has been argued that direct off-budget loans advanced under federal auspices, $14 billion in 1982, plus a portion of guaranteed loans (the total of which was $21 billion in 1982, and estimated at over $50 billion in 1983) represent unrecognized contributions to the true budget deficit.

Clearly, federal loans or loan guarantees which result in private expenditures which would not otherwise be undertaken must be recognized, aside from possible substitution effects, as augmenting aggregate demand. I should caution, however, against viewing them as, except possibly in small part, elements of a budget deficit in most, economically relevant senses. For Treasury borrowing to finance loans to the public, directly or indirectly, except as a second-order effect, adds neither to the net debt of the federal government nor the net financial assets of the public. (To the extent that the federal government borrows at the market rate of interest but lends at a below-market rate, it is, however, acquiring loan assets whose market value is less than the market value of its liabilities.) Hence (except for the parenthetical qualification), these off-budget federal borrowings do not have the essential wealth effects on consumption usually attributable to true budget deficits.

By another criterion as well, that of net claims on credit markets, federal borrowing to finance loans to the public would also qualify very partially, if at all, as components of a relevant budget deficit. For while, on the one hand, the Treasury borrowing draws on the public supply of credit, the Treasury or agency lending offers an essentially equivalent offset by meeting some of the public demand for credit.

II. Contingency Expenditures

Constructing a budget of "contingency expenditures" and associated receipts may well be a useful exercise. The relevant numbers, as we have suggested, are enormous, as may be the associated deficits (or surpluses). I should object, however, to incorporating measures of year-to-year increase in net contingent liabilities in current budget deficits.

Proclamations (usually intended to be alarmist) of huge net contingent liabilities and accruals to these net liabilities which might be considered annual deficits are in fact highly conjectural. What expenditures or outlays will finally be will depend frequently upon legislation of the future. Similarly, associated receipts will depend upon future legislation as well as highly uncertain economic and demographic developments. And how to balance projected expenditures and receipts depends further on uncertain and changing rates of discount.

An interesting and major case in point is the whole matter of Social Security, on which many have become exercised. We have been told that huge increases in net Social Security debt were having a major impact on aggregate consumption and saving. But after the initial sensationalist articles and arguments, a number of works (including my 1983 article) raised fundamental doubt as to the existence of any clear, measurable impact. And finally, in a few strokes of legislation, the balance between contingent expenditures and contingent liabilities was changed drastically. Who could even say accurately that the public over the last several decades has perceived annual increases in net Social Security wealth? And who can argue persuasively that they are perceiving such increases now?

III. Capital Budgets

Failure to separate current and capital expenditures in the formal federal budget and the associated failure to account systematically for capital assets contributes to confusion in economic analysis and consequent formulation of policy. It does make a difference whether a budget "deficit" finances transfer payments, current services, accumulation of stocks, or long-term investment. For example, the argument, to which I give little weight, that public perception of assets in the form of government debt is offset by anticipated associated tax liabilities, is negated to the extent that real income-producing government assets lie behind the debt. Federal capital accounts might of course well relate, further, not only to investment in physical assets owned by government, but to investment in the human capital and the private and public resources of the nation.

Objection to separating out capital expenditures from the current federal budget seems to stem considerably from concern that any resultant lowering of measured budget deficits would reduce resistance to excessive government spending. Aside from the ideological and unsubstantiated nature of the premise, however, there is no reason why, as I shall illustrate below, separate current and capital accounts cannot be combined to reach a bottom line showing precisely the unified budget deficit to which so much attention is now given. The separate subtotals, though, would give us a better clue as to the extent government was leaving a burden for future generations or building the houses—and other assets—that Sumner Slichter recommended for the young.

IV. Inflation Adjustments

The failure to account separately for capital expenditures quite breaks the publicly perceived connection between a budget deficit and spending beyond our means or squandering the public treasure. Recurring budget deficits have in fact been accompanied by *in*creasing federal net worth, as noted in my article with Pieper. What has also until recently been remarkably underplayed, if not ignored, is that in periods of inflation, our conventionally measured budget deficits are accompanied by *de*creasing values of real federal net debt. The "underlying reality...that every dollar of deficit...adds a dollar to debt" is simply not true in a real sense if price levels are not constant. And if interest rates fluctuate, the statement is not true even with reference to the market value of nominal debt. Adjustments of the official budget deficit to make it correspond to changes in net financial liabilities entail interest rate effects on market value of federal financial assets and liabilities and price or inflation effects on the real values of the corrected market values.

In general, we may write the adjusted or corrected deficit as

$$D_C = D - \dot{P}_A A + \dot{P}_B B - \dot{P}(B + M - A)$$

where $D =$ the "official deficit," $A =$ federal financial assets (excluding gold), $B =$ federal interest-bearing liabilities held by the public, $\dot{P}_A$ and $\dot{P}_B =$ the relative (weighted average) rates of price change of A and B, respectively, $\dot{P}$ = the inflation rate, $M =$ non-interest-bearing federal liabilities held by the public, essentially "high-powered money" or the monetary base, and $B + M - A =$ the "net debt."

As noted in my article with Pieper, the corrected deficit is vastly reduced and frequently converted to surplus in years of rapid inflation and rising interest rates (hence falling bond prices). The real federal net debt as defined above hence fell enormously since World War II, declining in per capita 1972 dollars from $3,694 at the end of 1946 to $1,445 by the end of 1980, despite a heavy preponderance of officially measured budget deficits in the intervening years.

What is more, turning to the high-employment deficit as a first-order measure of fiscal thrust, while the official measure of generally increasing deficit through the 1970's and up to 1981 suggested a stimulatory fiscal policy, the corrected measure showed quite the opposite. The severe recession of 1981 and 1982 was associated with high previous surpluses in the interest- and price-corrected high-employment budget—2.2, 1.8, 1.6, and 2.0 percent of *GNP* in the years 1978 to 1981. Interestingly, the recovery of 1983 is well accounted for by the sharp 1982 move to a deficit equal to 1.8 percent of *GNP* in the corrected high-employment budget; this was the product of a growing official deficit, slowing inflation and lower interest rates.

Recent work of Alex Cukierman and Jorgen Mortensen (1983) suggests that inflation corrections call for similar drastic revisions in perceptions of fiscal policies in Western Europe. In particular, the United Kingdom and Italy, with high official budget deficits, turn out to have had very low budget deficits, or even surpluses, after correction for inflation, perhaps accounting for their high unemployment rates. By contrast, the corrected budgets in West Germany, with relatively low rates of inflation and little or no net government debt, were sufficiently in deficit to have been stimulatory, possibly accounting for the low unemployment and rapid growth in Germany.

V. Equilibrium Growth

Whatever our measure of the deficit in dollars, pounds, lira, or marks, to what is it relevant in a large, changing, and generally growing economy? Recent work has suggested some apparent—if dubious—paradoxes involving deficits and rates of growth of interest-bearing securities and "money." We may get some needed perspective by exploring briefly one, simple overall relation. Netting out the acquisition of financial assets, essentially the off-budget items discussed above, we may write the official deficit as $D = \dot{B} + \dot{M} - \dot{A}$. Then, denoting $\gamma = (B + M - A)/Y =$ the ratio of net debt to *GNP*, and $g = \dot{Y}/Y =$ the rate of growth of *GNP*, it is readily apparent that $D/Y = \gamma g$ is the "equilibrium" condition for a stable ratio of net debt to *GNP*.

With the current value of γ in the United States approximately 0.3 and the growth of *GNP* in nominal terms about 10 percent, one notes that the "equilibrium" value of D/Y, far from zero, is some 0.03. Thus, at current or immediately anticipated rates of growth, we can accommodate an official deficit of 3 percent of *GNP* without increasing the relative debt burden. In fact, the deficit for fiscal 1983 ran closer to 6 percent of *GNP*, but it may be argued that this was largely a cyclical deficit and the structural or reasonably high-employment deficit would be considerably less.

Ignoring interest rate effects, we may write for our inflation-corrected, "equilibrium" budget deficit ratio, $D_C/Y = \gamma(g - \dot{P})$ or, more precisely, $D_C/Y = \gamma n$, where $n = (1+g)/(1+\dot{P})-1$. With $\gamma = 0.3$ and a real rate of growth, $n = 0.03$, an inflation-corrected budget deficit which would keep the ratio of net debt to output constant at its current value of 0.3 would then be in the order of 0.9 percent of *GNP*. This is in the neighborhood of the current price-adjusted high-employment deficit. Congressional Budget Office projections of an official high-employment deficit approaching 4 percent of *GNP* by 1988 would imply, on the basis of the concomitantly assumed 5 percent rate of inflation, however, a price-adjusted deficit ratio of some 2.5 percent. Such an *adjusted* deficit might well prove unsustainable. Relations estimated by Pieper and me suggest that resultant inflation and increases in debt would widen the gap between official and adjusted deficits until the latter were back at a more modest, equilibrium level.

VI. Conclusion—A Variety of Deficits

A few definitions and associated dollar figures for 1982 can help bring all this together. In Table 1, we note a variety of budgets or accounts to be considered. From them one can concoct a very large variety of deficits, only a subset of which are listed in Table 2. Some will be of greater relevance for certain purposes and some for others.

The current budget deficit, D1, would be most consistent with the more usual private business and state and local government measures, but the current adjusted deficit, D4, would be more economically relevant. In terms of macroeconomic policies for high-employment and stable price paths, my preference would go to D6, the national income account adjusted deficit, which would essentially indicate the change in net debt of the federal government to the public. As remarked above, in the years leading up to 1981, inflation and rising interest rates had both contributed to reducing national income adjusted deficits sharply or even converting some to surplus, but there was a marked shift in the correction in 1982.

We might also keep a close eye on D7, which would include price or inflation adjustments but not the (partially related) sharply fluctuating interest effects on the market value of debt. For a bottom line on the long-run trend in government activities, we might regard D8 and D12, the changes in net worth and "total net worth," the latter including contingent assets and liabilities. We should indeed, for many purposes go much further and recognize the intangible wealth to which government ostensibly contributes by expenditures for research and development, education and health.

What about D3, the official national income account budget deficit (or its unified budget twin)? We may as well continue to calculate it. We need the basic data in orderly form and we should have the series for continuity. But it may not merit prime focus. If we are concerned with economic analysis and consequences, we should not be shy about looking further.

TABLE 1—A VARIETY OF BUDGETS

Account (1)	Credits (2)	Debits (3)	Deficit[a] (4)
A. Current	617	744[b]	127
B. Capital	43[c]	63[d]	20
C. National Income	617	764	147
D. Net Revaluations of Financial Capital	−10	11	21
D*. Price Component	−28	−61	(34)
D**. Interest Component	18	72	55
E. Net Revaluations of Tangible Capital	9	–	(9)
F. Off-Budget (ΔV)	18	17	(1)
G. Contingent[e] (ΔV)	−390	−256	134

Note: Calendar 1982 unless otherwise indicated; ΔV is change in value of balances, and parentheses indicate surplus.

[a] Col. 4 is col. 3 minus col. 2.

[b] Includes capital consumption allowances and $701 billion in current outlays.

[c] Capital consumption allowances.

[d] Capital expenditures.

[e] Fiscal 1982.

TABLE 2—A VARIETY OF DEFICITS

Deficit Components[a] (Designation)	Deficit[b] (1)	 (2)
D1 = A (Current)	126.7	4.1
D2 = B (Capital)	20.4	0.7
D3 = $A+B=C$ (National Income)	147.1	4.8
D4 = $A+D$ (Current Adjusted)	147.8	4.8
D5 = $A+D^*$ (Current Prices-Adjusted)	93.1	3.0
D6 = $C+D$ (*NI* Adjusted)	168.1	5.5
D7 = $C+D^*$ (*NI* Prices-Adjusted)	113.4	3.7
D8 = $A+D+E$ (Net Worth)	139.1	4.5
D9 = $C+F$ (*NI* plus Off-Budget)	146.1	4.8
D10 = $C+F+G$ (Total)	280.1	9.1
D11 = $C+D+F+G$ (Total Adjusted)	301.1	9.8
D12 = $A+D+E+F+G$ (Total Net Worth)	272.1	8.9

Note: Calendar 1982, except for D10, D11, and D12, which includes the G component for fiscal 1982; *NI* = National Income.

[a] As designated in Table 1.

[b] Col. 1 is billions of dollars; col. 2 is percent of *GNP*.

REFERENCES

Cukierman, Alex and Mortensen, Jorgen, "Monetary Assets and Inflation Induced Distortions of the National Accounts–Conceptual Issues and Correction of Sectoral Income Flows in 5 EEC Countries," *Economic Papers* No. 15, June 1983, Commission of the European Communities, Directorate-General for Economic and Financial Affairs, Internal Paper.

Eisner, Robert, "Social Security, Saving, and Macroeconomics," *Journal of Macroeconomics*, Winter 1983, *5*, 1–19.

______ **and Pieper, Paul J.,** "A New View of the Federal Debt and Budget Deficits," *American Economic Review*, March 1984, *74*, 11–29.

Public Opinion and the Balanced Budget

By ALAN S. BLINDER AND DOUGLAS HOLTZ-EAKIN*

Like wage-price controls, balancing the federal government budget has long enjoyed greater popularity with the public than with economists. A poll taken after a decade of the Great Depression, for example, showed that 61 percent of the populace was willing to cut federal spending immediately by enough to balance the budget, while only 17 percent were opposed. (See Herbert Stein, 1969, p. 118.) Nor has this idea's popularity declined over time.

In 1949, only 38 percent felt that the government should incur a deficit "to avoid the possibility of another depression." In 1953, 69 percent favored balancing the budget over a personal tax cut. In a similar vein, 79 percent of those surveyed in 1979 supported a tax cut, but only 38 percent continued to favor lower taxes if it imposed a larger federal budget deficit. (All survey data are from *The Gallup Poll.*)

Why the continued popularity of this proposal? One possibility is that the polling data reflect simple-minded homilies about the evils of debt based on invalid analogies to personal finances ("Neither a borrower nor a lender be"). A second somewhat related possibility is that the attractiveness of balanced budgets reflects a general ideological attachment to fiscal conservatism. But these are not the only possibilities. It could be that support for balancing the budget is based on more or less coherent beliefs about how the economy and/or the government works. For example, some economists have favored a balanced budget as a way to control federal spending. Indeed, *The Harris Survey* in 1982 found that Americans agreed by more than a 2 to 1 margin that a constitutional mandate to balance the federal budget would be "an effective way to keep federal spending under control."

While most Americans favor a *mandatory* balanced federal budget, not all do. This paper uses cross-sectional differences among respondents to two public opinion polls to try to discriminate among competing hypotheses about why Americans want the budget balanced. In Section I the data and statistical methods are briefly described. Sections II and III present the results of analyzing data from two different public opinion polls taken at about the same time.

I. Data and Methods

The data used are individual responses to a *Gallup Poll* and a *CBS/New York Times Poll* conducted in March and April of 1980, respectively, a time when the proposed balanced budget amendment to the Constitution was very much in the news and there was great public concern about inflation. The *Gallup Poll* asked about support for the amendment. The *CBS/NY Times Poll* asked respondents if they favored a requirement for a balanced budget even if it would require cutbacks in federal spending.

In each case, a large majority favored a balanced budget requirement. In the *Gallup Poll*, the margin was 67 percent in favor, 13 percent opposed, and the rest undecided; thus, among those expressing an opinion, an astounding 84 percent favored the amendment. In the *CBS/NY Times Poll*, which stressed the need for cutbacks in spending, 61 percent supported budget balance while 32 percent were opposed (only 7 percent were undecided). The difference between the two sets of answers to apparently similar questions is striking testimony to the sensitivity of polling results to the precise wording of the question. It also suggests an

*Princeton University, Princeton, NJ 08544. Data were kindly provided by the Roper Center and the Inter-University Consortium for Political and Social Research. We thank the Social Science User's Service at Princeton University and the Institute for International Economic Studies for assistance, Michael Kagay of Louis Harris for providing polling results, and Assar Lindbeck, Soren Blomquist, and John Helliwell for helpful comments.

unwillingness to face the real costs of balancing the budget. In fact, a 1981 *Harris Survey* found that in no instance were a majority of respondents willing to reduce spending on any domestic program rather than unbalance the federal budget.

Both polls include the standard socioeconomic characteristics of respondents such as age, sex, race, education, political affiliation, and income. Beyond this, each poll has a particular strength. In the *Gallup Poll*, respondents were asked to present both the best argument in favor of requiring a balanced budget and the best argument against it. They were also asked how they thought balancing the budget would affect the rate of inflation, tax burdens, federal government employment, and federal government spending. Their answers give us an interesting glimpse of individuals practicing amateur economic analysis.

The *CBS/NY Times Poll* inquired more closely into political ideology, economic values, and personal economic circumstances. Respondents were asked whether they felt inflation or unemployment was the greater economic problem, whether they would accept greater unemployment in order to reduce inflation, and what they thought was the best way to fight inflation.

Though a logit model was fit to the results of each poll, differences in the data dictated rather different specifications of the independent variables. In each case, however, there were so many potential right-hand variables that some preliminary data screening was necessary.[1] One major result of this process was to eliminate all the people who answered "don't know" to the balanced budget question, as there appeared to be no significant information on these people.

In an effort to "play fair," the equations were first estimated by entering all potential independent variables on the right-hand side. Variables were then eliminated on the basis of *t*-tests conducted at the 10 percent level. Once a final specification was arrived at, previously eliminated variables were reentered and tested for significance.

II. Results from the *CBS/NY Times Poll*

The *CBS/NY Times Poll* asked the following question:

> To deal with our economic problems, would you favor or oppose requiring a balanced budget even if it means spending less on military and domestic programs?

A logit model fit to the responses (yes = 1) gave the results presented in Table 1. Estimates are based on 1,262 cases remaining after initial screening for "don't knows" and missing data.

TABLE 1—LOGIT ESTIMATES FROM *CBS/NY TIMES POLL*

Variable	Coefficient[g]	Partial Effect[f]	Sample Mean
Socioeconomic Characteristics			
Black	.59 (2.5)	.12	.090
Male	−.37 (−3.0)	−.08	.477
College Graduate	.38 (2.4)	.08	.230
Personal Economic Circumstances			
Recent Layoff[a]	.35 (1.9)	.07	.165
Better Off[b]	−.28 (−1.7)	−.06	.190
Attitudes and Values			
Best Way to Fight Inflation[c]	.86 (6.9)	.19	.526
Inflation Worse than Unemployment[d]	.23 (1.7)	.05	.802
Willing to Tradeoff[e]	.27 (1.9)	.06	.269
Constant	.056 (0.3)	–	–
Likelihood Ratio Statistic = 81.6			
Likelihood Ratio Index[f] = .051			

[a] Laid off during last year.
[b] Better off than last year.
[c] Balancing the budget is the best way to fight inflation.
[d] Inflation is a greater economic problem than unemployment.
[e] Willing to let unemployment rise to fight inflation.
[f] See text for explanation.
[g] *t*-statistics are shown in parentheses.

[1] Details of this screening process are included in Holtz-Eakin (1983).

Since many of the important explanatory variables are attitudinal, and hence themselves cry out for explanation, we estimated the following "decision tree."[2] Socioeconomic characteristics and personal economic circumstances are assumed to influence ideology and value judgments. All of these variables, in turn, are assumed to influence views on how the economy works (for example, are deficits inflationary?). Finally, personal characteristics, ideology, and economic judgments are all assumed to influence attitudes toward balancing the budget. To keep the task manageable, we worked "down" the decision tree. First, the determinants of the balanced budget choice were estimated —including economic judgments, ideological factors, etc. Then, the determinants of these, lower level, decisions were estimated. The process continued until the decision tree (and researchers) were exhausted.

Some explanation of Table 1 is in order. The first column gives the estimated logit coefficient and its *t*-statistic (shown in parentheses). However, the importance of each variable is best understood by the partial effect shown in the next column. This measure is computed as the change in the probability of answering "yes" to the balanced budget question as each dummy variable is varied from zero to one, holding all other variables at their sample means.

For example, believing that balancing the budget is the best way to fight inflation increases the probability of favoring budget balance by .19. This is far and away the most important determinant of the dependent variable, reflecting the overwhelming concern about inflation at the time of the poll. (The other choices offered for how to fight inflation were cutting taxes, imposing wage-price controls, or none of the above.)

Among the socioeconomic variables, only race, sex, and education had any significant impact on attitudes toward balancing the budget. Blacks were .12 *more* likely to support a balanced budget requirement, men were .08 less likely, and those with a college education or more were .08 more likely. The result for blacks is certainly surprising, given their typical position on the economic ladder.

At least as interesting as this short list of variables that mattered is the much longer list of obvious socioeconomic variables that apparently, and often surprisingly, have no bearing on support for the balanced budget. These include age, income, political ideology, and party affiliation.

Variables reflecting personal economic circumstances also held some surprises: those who have been laid off within the past year were .07 *more* likely to support the balanced budget, while those who are better off then they were a year ago were .06 *less* likely. Perhaps people who are doing well want things left alone while those who are doing poorly seek change.

Two other attitudinal variables were found to be of some importance. Those who felt that inflation is a greater economic problem than unemployment were .05 more likely to favor a balanced budget and those who were willing to use unemployment to fight inflation were .06 more likely. These are appealing results. What is striking is the strength of the anti-inflation sentiment: 80 percent of the sample felt that inflation was the greater problem, an unusually high proportion. (Compare Douglas Hibbs, 1982.) Note, however, that only 27 percent of the sample was willing to trade higher unemployment for lower inflation. Thus, while anti-inflation sentiment was high, willingness to bear the burden of contractionary policy was not.

In keeping with our decision tree model, logit equations were also estimated for each attitudinal variable. We have space only to describe briefly the most important of these: the equation explaining whether or not the respondent thought that balancing the budget was the best way to fight inflation. The items of major importance were as follows. Individuals with family incomes in excess of $10,000 (in 1979), and those who felt there was a weak link between budget deficits and inflation, were less likely to come to this conclusion. Those willing to use unemployment to fight inflation, and conservatives, were more likely to reach this conclusion. Thus, while ideology shows no *direct* effect on the balanced budget question, there is an *indirect* effect. Finally, respondents with col-

[2] For a more detailed explanation of the model, including a justification of the logit specification based on utility maximization, see Holtz-Eakin.

lege or greater education tended to conclude that balancing the budget was a good anti-inflation strategy. Notice that this effect of higher education reinforces the tendency found above—that college graduates tend to be more in favor of balancing the budget.

Finally, Table 1 offers two measures of goodness of fit. The first is the standard likelihood ratio statistic, $-2\log(LS/LC)$, where LC is the likelihood at convergence and LS is the likelihood when each case is assigned a probability of answering yes equal to the sample frequency. The second is a likelihood ratio index defined as $(\log LC/\log LS)-1$. This measure gives the percentage improvement in the log-likelihood due to using individual data. This goodness of fit statistic is only 0.051, corresponding roughly to an *OLS* R^2 of 0.062. Thus our ability to predict attitudes toward balancing the budget from the information available in the *CBS/NY Times* poll is meager.

III. Results from *the Gallup Poll*

The Gallup Poll was simpler to handle. After screening out "don't knows" and missing data, 1,260 observations were left and a single logit equation was estimated to explain the answers to the following question (yes = 1):

> A proposed amendment to the Constitution would require Congress to approve a balanced federal budget each year. Government spending would have to be limited to no more than expected revenues, unless a three-fifths majority of Congress voted to spend more than expected revenue. Would you favor or oppose this amendment to the Constitution?

Results are displayed in Table 2. The most interesting results pertain to the arguments for and against the amendment.

TABLE 2—LOGIT ESTIMATES FROM *THE GALLUP POLL*

Variable	Coefficient[f]	Partial Effect	Sample Mean
Socioeconomic Characteristics			
Age[a]	−0.172 (−3.1)	−.014	44.2
Democrat	−.40 (−2.1)	−.03	.411
Union[b]	.39 (1.6)	.03	.244
Full-Time Student	−1.46 (3.1)	−.20	.030
High School[c]	.57 (2.7)	.045	.394
Graduate School[d]	−.96 (3.0)	−.11	.066
Arguments in Favor of Amendment			
Live within Means[e]	1.73 (6.2)	.11	.254
To Fight Inflation	1.75 (6.1)	.105	.241
Will Reduce Wasteful Programs	2.12 (7.1)	.13	.274
Miscellaneous	1.23 (3.1)	.066	.060
There is None	−1.57 (−3.7)	−.22	.034
Arguments Against Amendment			
Will Hurt the Economy	−1.09 (−3.9)	−.12	.127
Too Restrictive	−1.74 (−7.5)	−.12	.196
Will Reduce Necessary Programs	−1.225 (−4.4)	−.14	.146
Miscellaneous	−1.97 (−5.2)	−.31	.053
Perceived Effects of Amendment			
Small Effect (+ or −) on Inflation	.83 (2.1)	.08	.769
Lower Inflation a Lot	1.35 (2.9)	.08	.191
Raise Taxes a Lot	−.67 (−2.6)	−.07	.114
Lower Welfare Spending a Lot	−.65 (−3.0)	−.06	.213
Constant	1.65 (3.3)		
Likelihood Ratio Statistic = 302.4			
Likelihood Ratio Index = .276			

[a] This is a continuous variable measured in years. The partial effect refers to raising age from 39.2 to 49.2.

[b] Respondent or spouse is a union member.

[c] High school or technical school education.

[d] Some postcollege education.

[e] Best argument is that everyone should live within his means, or that the national debt is already too high.

[f] t-statistics are shown in parentheses.

Respondents divided almost evenly among three general arguments in favor: that nations (like people) should "live within their means," that balancing the budget is anti-inflationary, and that balancing the budget is a good way to cut wasteful government programs. Certainly, most economists would agree that there are important grains of truth in the latter two arguments, while the first seems to reflect the naive homilies mentioned at the outset.[3] People selecting any of these three arguments are about .10 to .13 more likely to support the amendment. Believers in crowding out are presumably included in "miscellaneous."

The most popular argument against the amendment (selected by about 20 percent of the sample) is that it would tie the hands of policymakers. Others worried that it would reduce necessary military and domestic programs (15 percent),[4] or hurt the economy in times of emergency, that is, interfere with stabilization policy (13 percent). Those who selected one of these three arguments were .12 to .14 less likely to support the amendment.

It is also fascinating to note that 77 percent of the respondents thought that a balanced budget amendment would have only a small effect on inflation (up or down), while 19 percent thought it would lower inflation substantially. (Recall that 53 percent of the respondents to the *CBS/NY Times Poll* nonetheless believed that balancing the budget is the best way to fight inflation.) Either of these groups was .08 more likely to support the amendment than the small minority who thought budget balancing was strongly inflationary. Finally, the minorities (11 and 21 percent, respectively) who thought that a balanced budget amendment would lead to large increases in taxes or to large cuts in welfare spending were .06 to .07 less likely to support the amendment. The latter result surprised us.

A few socioeconomic variables were also significant. Full-time students (only 3 percent of the sample) were much more opposed to the amendment, and those with some education beyond college (7 percent of the sample) were moderately more opposed. These results on education contradict those obtained in the *CBS/NY Times Poll.* Older people and Democrats were less in favor of the amendment, and union members were more in favor; but each of these effects is small.

As in the case of the *CBS/NY Times Poll*, it is just as interesting to note that many socioeconomic variables typically thought of as important determinants of opinions toward federal budget policy were not significant. These include race and sex, which turned out to matter in the *CBS/NY Times Poll*, and income and geographical region, which did not.

Finally, notice that the likelihood ratio index of goodness of fit is 0.276, a five-fold improvement over the *CBS/NY Times Poll*, and a figure which seems quite respectable relative to standard econometric results based on cross sections of individuals. (It corresponds roughly to a *OLS* R^2 of .25.) One of the most encouraging aspects of the results is the importance of economic reasoning in obtaining this fit. The presumed effects of the balanced budget amendment on inflation, taxes, and welfare programs all impact on the decision. In addition, many individuals cite economic arguments for and against the amendment, and these arguments affect their opinions.

IV. Conclusion

What may we conclude from this exercise? Clearly, Americans favor some sort of balanced budget restriction, and probably always have. However, they favor it for a smorgasbord of reasons and at an unclear price.

From an economist's perspective, it is encouraging that political affiliation, ideology, and personal circumstances matter far less than economic rationales. The best evidence for this is the vastly superior fit of the *Gallup Poll* estimates, which relied on information about respondents' economic reasoning, over the *CBS/NY Times Poll* estimates, which relied more on individual characteristics and ideology. If this correlation really signifies causation, rather than rationalization of a

[3] This argument actually aggregates many similar responses such as "the federal budget is no different than my budget," "You can't keep spending more than you take in," etc.

[4] The schizophrenic attitude of Americans toward government spending is evident. Nobody wants to undertake actions that will eliminate "necessary" programs, but using a blunt instrument like the balanced budget amendment to fight "waste" is perfectly acceptable.

decision reached on ideological grounds, then rational public discourse on government budget deficits may one day be possible.

REFERENCES

Harris, Louis, *The Harris Survey*, No. 81, 1981, released October 8, 1981; No. 69, 1982, released August 30, 1982.

Hibbs, Douglas, "Public Concern About Inflation and Unemployment in the United States: Trends, Correlates, and Political Implications," in Robert Hall, ed., *Inflation: Causes and Effects*, Chicago: University of Chicago Press, 1982.

Holtz-Eakin, Douglas, "An Analysis of Public Preferences for Balanced Federal Budgets," unpublished, Princeton University, 1983.

Stein, Herbert, *The Fiscal Revolution in America*, Chicago: University of Chicago Press, 1969.

The Gallup Poll, Computer tape, Roper Center, March 1980.

Public Opinion, various years.

CBS/New York Times Poll, Computer tape, Inter-University Consortium for Political and Social Research, April 1980.

Wage-Price Behavior in the National Models of Project LINK

By BERT G. HICKMAN AND LAWRENCE R. KLEIN*

The purpose of this paper is to compare the wage and price dynamics in some major OECD countries, with particular reference to the relationship between unemployment and inflation and to nominal and real wage flexibility under demand and supply shocks. On other occasions the wage equations and the price equations have been separately studied within the LINK system. This study extends that work in new directions, mainly through techniques of model simulation.

If government spending is fixed at two different levels for a given period, with all other exogenous variables constant, we get two sets of reduced-form values of the inflation and unemployment rates by dynamically simulating the LINK system. It is our hypothesis that the result will be a negative or tradeoff relationship between inflation and unemployment. If the price of oil is varied, on the other hand, we expect a positive or stagflationary relationship between unemployment and inflation.

It is a frequent mistake to confound the tradeoff relationship with the wage determination equation. The latter is a structural equation associating wage changes with labor market conditions, say, the imbalance between supply and demand, represented by unemployment. The tradeoff between inflation and unemployment is a proper subject for analysis, but not in the context of structural equations, only in the context of pairs of reduced-form solutions. The structural Phillips curve could be fairly stable, while the relationship between unemployment and inflation shifts because of supply shocks or changes in productivity behavior.

The core relationships representing the wage-price structure in the LINK models are the labor market Phillips curve (except in the U.K. model) and the price markup equation. Thus wage rate change is a function of price change and unemployment, whereas prices are set as a markup over significant unit costs, namely, import costs and normal unit labor costs, which depend in turn on the wage rate and productivity.

I. Design of the Simulations

Results are presented for the following LINK models: Canada (FOCUS), France (METRIC), Germany (SYSIFO), Japan (KIER), the United Kingdom (LBS), and the United States (WEFA). The demand shock scenarios are for a sustained increase in real government spending equal to 1 percent of *GNP* in each country in turn, with induced international interactions accounted for in this open-economy linked system with endogenous exchange rates and trade flows. The supply shock imposes a step increase of 30 percent in the import price of oil, impinging simultaneously on all countries. Neither shock is accommodated by monetary policy, although the discount rate in the Japanese model reacts endogenously to changes in the foreign interest rate, the exchange rate, and the inflation rate, and endogenous changes occur in the money supply for given reserves in the other countries. These are dynamic simulations based on a control solution for 1984–88.

II. Results for the Fiscal Policy Scenario

The dynamic responses to the fiscal shock, as calculated in the first, third and fifth years to adjustment, are shown in Table 1 for real *GNP* (Y), employment (L), the unemploy-

*Stanford University, Stanford CA 94305, and University of Pennsylvania, Philadelphia, PA 19104, respectively. We are indebted to C. A. Bolino, Michael Papaioannou, and Christian Peterson for very helpful assistance on this paper.

TABLE 1—DYNAMIC RESPONSES TO FISCAL SHOCK

Model	*Y*	*L*	*U*	*W*	*PC*	*DW*	*DP*
			First Year				
CA	0.8	0.4	−0.2	0.3	−0.1	0.3	−0.1
FR	1.4	0.1	−0.1	0.0	0.0	0.0	0.0
GE	1.1	0.2	−0.2	0.1	0.0	0.1	0.0
JA	1.3	0.1	−0.1	0.5	0.1	0.5	0.1
UK	0.8	2.0	−1.7	0.1	0.3	0.1	0.3
US	1.1	0.3	−0.3	0.0	0.0	0.0	0.0
			Third Year				
CA	0.8	0.7	−0.4	1.3	0.7	0.6	0.3
FR	1.1	0.2	−0.2	1.0	0.9	0.6	0.5
GE	1.1	0.5	−0.5	1.3	0.5	0.7	0.3
JA	1.6	0.1	−0.1	2.7	0.9	1.2	0.5
UK	1.1	0.9	−0.8	1.5	0.3	1.3	0.4
US	0.6	0.7	−0.7	0.8	0.8	0.5	0.5
			Fifth Year				
CA	0.5	0.8	−0.5	2.5	1.4	0.6	0.4
FR	0.9	0.2	−0.2	1.7	1.5	0.3	0.2
GE	1.2	0.5	−0.5	2.9	1.2	0.8	0.4
JA	1.5	0.0	0.0	3.6	1.4	0.0	0.1
UK	0.8	0.8	−0.7	3.3	1.8	0.8	0.7
US	−0.1	0.3	−0.3	1.7	1.4	0.4	0.2

ment rate (U), the nominal wage level (W) and inflation rate (DW), and the consumer price level (PC) and inflation rate (DPC). The responses are measured either by the percentage differences between the shock and baseline levels (Y, L, W, PC), or by the absolute change in the percentage rates of unemployment and inflation in the two solutions (U, DW, DPC).

The impact multipliers for output show increases ranging between 0.8 and 1.4. The accompanying unemployment responses are rather small for all countries except the United Kingdom, although the decrease for Japan is from a base rate on only 2 percent, as compared with about 10 percent elsewhere. With the exception of the United Kingdom, fluctuations in output are being accommodated by large changes in hours and productivity, owing to extensive labor hoarding, so that employment and unemployment change relatively little. In this initial year as later, changes in unemployment mirror those of employment, since induced changes in labor supply are insignificant.

The induced wage and price movements are also small at first, partly because unemployment changes so little, but also owing to lags in the response of wages to unemployment. In the U.K. model, the wage rate is not directly related to unemployment and it responds to output with a substantial lag, so that the nominal wage scarcely moves in the first year despite the pronounced drop in unemployment.

Wage pressures build up considerably by the third year of adjustment, partly because unemployment falls further in many models, but also because of the distributed lags in the response of wages to unemployment, prices to wages, and wages to prices in the structural equations. Output and employment stabilize or fall between the third and fifth years in most models, but wages and prices increase further as the lags work their way through the system.

The expected negative tradeoff between unemployment and the rate of consumer price inflation is found for all countries. Prices are largely undisturbed as unemployment falls in the first year, but the tradeoff becomes less favorable over time. The tradeoff curve for unemployment, as evaluated in the third year, is steeper for France and Japan than in the other countries, although it is important to note that this is not true of the tradeoff between output and inflation. The price level continues its rise in all countries after the third year, but the inflation rate decelerates or declines except in the United Kingdom, as expected under a nonaccommodating monetary policy.

III. Stagflation under an Oil Shock

An adverse oil shock leads to the anticipated stagflationary results (Table 2). Real *GNP* rises by small amounts in the first year in Canada, France, and Japan, as real imports are curtailed by price increases and export demand is stimulated by OPEC respending, but even these countries experience production declines in later years. Nominal wages and prices rise in all models through the third year of adjustment, although the inflation rate is already subsiding by that year for Canada, Germany, Japan and the United States, as expected for a once-and-for-all price shock without monetary accommodation. The increased unemployment from the oil shock builds up grad-

TABLE 2—DYNAMIC RESPONSES TO OIL PRICE SHOCK

Model	*Y*	*L*	*U*	*W*	*PC*	*DW*	*DPC*
			First Year				
CA	0.2	0:0	0.0	0.2	0.3	0.2	0.3
FR	0.5	0.1	−0.1	0.6	0.7	0.6	0.7
GE	−0.3	−0.1	0.1	0.3	0.6	0.3	0.6
JA	0.2	0.0	−0.0	0.5	0.7	0.5	0.7
UK	−0.1	−0.0	0.0	0.6	0.7	0.6	0.7
US	−0.9	−0.3	0.4	0.8	1.4	0.8	1.4
			Third Year				
CA	−0.1	−0.1	0.1	0.4	0.5	0.1	0.1
FR	−0.1	0.1	−0.1	3.7	3.9	1.8	1.7
GE	−0.8	−0.5	0.5	1.2	1.6	0.3	0.4
JA	−0.3	−0.1	0.0	1.2	1.3	0.1	0.1
UK	−1.0	−0.0	0.1	1.5	2.2	0.2	0.7
US	−1.4	−1.0	0.9	1.5	1.7	0.1	0.1
			Fifth Year				
CA	0.1	0.0	−0.0	0.7	0.8	0.2	0.1
FR	−1.4	−0.1	0.1	6.9	7.2	1.6	1.7
GE	−0.6	−0.6	0.6	1.1	1.9	−0.1	0.1
JA	−0.4	−0.1	0.0	0.3	1.0	−0.5	−0.2
UK	−1.1	0.0	−0.0	0.1	1.9	−0.8	−0.4
US	−1.0	−1.0	0.7	1.0	1.3	−0.3	−0.3

TABLE 3—WAGE AND PRICE RESPONSES TO FISCAL SHOCK

Model	*W*	*W/P*	*PC/P*	*W/PC*	*PX/PM*	*Y/L*
			First Year			
CA	0.3	0.0	−0.4	0.4	0.7	0.4
FR	0.0	0.1	0.1	0.0	−0.3	1.3
GE	0.1	0.0	−0.1	0.1	0.0	0.9
JA	0.5	0.3	−0.1	0.4	0.1	1.2
UK	0.1	−0.2	0.0	−0.2	−0.2	−1.2
US	0.0	−0.1	−0.1	0.0	0.0	0.8
			Third Year			
CA	1.3	0.2	−0.4	0.6	0.5	0.1
FR	1.0	0.0	−0.1	0.1	−0.2	0.9
GE	1.3	0.5	−0.3	0.8	0.1	0.6
JA	2.7	1.5	−0.3	1.8	0.1	1.5
UK	1.5	1.2	0.0	1.2	−0.2	0.2
US	0.8	0.0	0.0	0.0	0.1	−0.1
			Fifth Year			
CA	2.5	0.6	−0.5	1.1	0.4	−0.3
FR	1.7	0.1	−0.1	0.2	0.6	0.7
GE	2.9	1.2	−0.5	1.7	0.4	0.7
JA	3.6	1.8	−0.4	2.2	0.5	1.5
UK	3.3	1.3	−0.2	1.5	0.1	0.0
US	1.7	0.3	0.0	0.3	0.3	−0.4

ually, and after the third year, wages, and hence prices, are depressed in Japan, the United Kingdom, and the United States, and the inflation rate levels off or declines in Canada, France, and Germany.

The movements of nominal and real wage rates under the two shocks will be examined next. Are nominal wages substantially stickier in the United States than in other OECD countries, as contended by Jeffrey Sachs (1979, 1983), William Branson and Julio Rotemberg (1980), and Robert Gordon (1982)? Are real wages more flexible in the United States than in Japan and Europe, as argued by Sachs and Branson-Rotemberg, but questioned by Gordon? To our knowledge these questions have not previously been addressed in the context of complete macroeconometric models allowing for domestic and international interactions among product, labor, and financial markets under uniform demand and supply shocks. This is important because the behavior of nominal and real wages in response to shocks is strongly affected by conditions outside the labor market itself and should be studied in a general equilibrium framework.

IV. Wage Behavior under a Fiscal Shock

The results summarized in Table 3 indicate that nominal wage levels are somewhat stickier under a demand shock in the United States, but not strikingly so or in comparison to all other countries. The increases for France and the United States are about the same over the entire simulation horizon. The initial increases are small to moderate for Canada, Germany, and the United Kingdom, but they build up substantially over time. The Japanese model shows the fastest and largest increases of all.

With regard to real wages under a demand shock, they are quite rigid in France and the United States, whereas sizable gains are found in the other models after the first year. Again, the largest rise occurs in Japan, but substantial increases are found also for Germany and the United Kingdom. As a rule, consumer prices (*PC*) rise somewhat less than the *GNP* deflator (*P*), so that the real consumption wage rises more than the real product wage in all countries except the United Kingdom and the United States by the third year, and the United States by the fifth. The *GNP* deflator rises more than consumer prices in part because the demand shock raises the prices of exports (*PX*) more than imports (*PM*).

The rigidity of the real wage in the French model is explicable by quick and complete indexation of wages to prices. Such prompt indexation to lagged inflation is not characteristic of the U.S. economy, however. Rather, another factor for the United States is the inclusion in the WEFA model of interest rates in the price equations, so that prices are set on the basis of costs of capital as well as labor and materials. Since interest rates rise for a nonaccommodated fiscal shock, a direct interaction is established between prices and wages to quicken the adjustment to full long-term indexation.

Average productivity in terms of output per employee rises in the short run with demand expansion in all models except the U.K. model. By the third year only France, Germany, and Japan still record substantial gains. The changes in real product wages and labor productivity are about equal in that year for Canada, Germany, Japan, and the United States. The productivity gain exceeds the increase in the real product wage in France, however, reducing the share of labor income, whereas the reverse is true of the United Kingdom.

TABLE 4—WAGE AND PRICE RESPONSES TO OIL PRICE SHOCK

Model	*W*	*W/P*	*PC/P*	*W/PC*	*PX/PM*	*Y/L*
			First Year			
CA	0.2	0.1	0.2	−0.1	−1.1	0.2
FR	0.6	1.7	1.8	−0.1	−5.9	0.4
GE	0.3	0.1	0.4	−0.3	−5.7	−0.4
JA	0.5	0.8	1.0	−0.2	−12.9	0.2
UK	0.6	0.2	0.3	−0.1	0.1	−0.1
US	0.8	−0.1	0.5	−0.6	−4.7	−0.6
			Third Year			
CA	0.4	0.2	0.3	−0.1	−1.5	0.0
FR	3.7	1.3	1.5	−0.2	−5.8	−0.2
GE	1.2	0.1	0.5	−0.4	−6.2	−0.3
JA	1.2	1.1	1.2	−0.1	−16.2	−0.2
UK	1.5	−0.5	0.2	−0.7	0.4	−1.0
US	1.5	0.2	0.4	−0.2	−4.5	−0.4
			Fifth Year			
CA	0.7	0.2	0.3	−0.1	−1.1	0.1
FR	6.9	0.6	0.9	−0.3	−3.4	−1.3
GE	1.1	−0.1	0.7	−0.8	−6.2	0.0
JA	0.3	0.6	1.3	−0.7	−16.1	−0.3
UK	0.1	−1.6	0.2	−1.8	0.2	−1.1
US	1.0	0.2	0.5	−0.3	−5.1	0.0

V. Wage Responses to an Oil Shock

Finally, in Table 4, we summarize the effects of a supply shock. Interestingly enough, nominal wages rise more under the initial impact of an oil price increase in the United States than in the other countries. On the other hand, the real consumption wage falls furthest in the U.S. simulation during the first year of adjustment. Thus it is nominal price flexibility, rather than nominal wage rigidity, which accounts for the sizable real wage decline in the U.S. model under an adverse energy shock, contrary to the usual argument. The relevant category of price flexibility, however, is with respect to increases in energy prices rather than to changes in nominal wages. Moreover, the same structural characteristic of the U.S. model which results in a small nominal wage response to a demand shock—a flat labor market Phillips curve—makes for a large nominal wage response to the rise of consumer prices under an energy shock. This is because the increased unemployment from the energy shock does not put much downward pressure on wage bargains to counteract the upward pressure stemming from the increased consumer prices.

Nominal wages increase further between the first and third years in all countries. The buildup is especially large in the French model, since real activity rises at first under the oil shock and adds to the wage pressures from higher consumer prices. Except for France, wage pressures moderate in all countries between the third and fifth years, with substantial reductions in wage levels occurring in the Japanese, U.K., and U.S. models.

Real consumption wages are depressed in all models and years as a result of the oil shock, but the real product wage is increased at least slightly in most countries, including the United States after the first year. As emphasized by Sachs and others, the deterioration in the terms of trade as a result of an oil shock tends to increase the consumer price level, which includes import costs, relative to the *GNP* deflator, which excludes them. The wedge between the real consumption and product wages is small for Canada and the United Kingdom, however, owing to the energy resources which insulate them from an adverse terms of trade effect, though not from the adverse price level and productivity effects of an energy shock.

The relative movements of the real product wage and labor productivity imply increased labor shares of value-added in all countries, with the exception of the final year in the United Kingdom. Again as stressed by Sachs, real consumption wages would need to fall more than labor productivity to preserve the profit share.

Investment expenditure is also depressed by an oil shock in all the models, but this is not a consequence of the squeeze on the profit share, as under Sachs' hypothesis. Rather, investment depends on real income or output and on the rental price of capital in these models. In the case of an energy shock, the initial downward impetus to real consumption from the induced reduction in the real wage and, in the Japanese, German, and U.K. models, also from negative real balance or wealth effects, itself depresses investment demand through an accelerator effect. Interest rates also rise to restrict investment because the shock is not accommodated by easier money in this scenario.

VI. Macroeconomic Equilibrating Mechanisms

In conclusion, we wish to emphasize that equilibrating forces are at work in these models, albeit they act slowly because of long adjustment lags. These are not neoclassical mechanisms, however. Real consumption wages rise with an expansionary fiscal policy and fall with an adverse oil shock, but in neither case is labor supply much affected. The real product wage rises in both scenarios in most of the models, but the increase is insufficient to reduce labor demand to the preshock level in the fiscal case, and it is in the wrong direction for equilibration of employment under the energy shock. Nevertheless, the dynamic simulations show a long-run tendency for output and employment to revert toward the control solution in most of the models under both shocks. This is because of the familiar macroeconomic adjustment mechanisms involving induced changes in the price level, namely, the Keynes and Pigou effects on interest rates and saving decisions.

Under an expansionary fiscal policy without monetary accommodation, inflation erodes real money balances and interest rates rise to crowd out gradually the interest-sensitive components of investment and consumer demand. Additionally, consumption demand may be directly reduced by adverse real balance or wealth effects in some of the models.

In the case of an adverse oil shock, the induced rise of unemployment will tend eventually to reverse the upward movement of wages and prices and cause real money balances and real wealth to increase once more. This will augment investment and consumption demand to restore that portion of the real income decline due to transitory demand deficiencies, as distinguished from the permanent loss of productivity and potential output from the oil shock. An expansionary monetary policy could accomplish this task with much less disruption to production and employment, however, at the cost of a permanently higher price level, though not a permanent increase in the inflation rate.

REFERENCES

Branson, William and Rotemberg, Julio, "International Adjustment with Wage Ridigity," *European Economic Review*, May 1980, *13*, 309–32.

Gordon, Robert J., "Why U.S. Wage and Employment Behavior Differs From That in Britain and Japan," *Economic Journal*, March 1982, *92*, 13–44.

Sachs, Jeffrey, "Wages, Profits, and Macroeconomic Adjustment: A Comparative Study," *Brookings Papers on Economic Activity*, 2:1979, 269–319.

______, "Real Wages and Unemployment in OECD Countries," *Brookings Papers on Economic Activity*, 1:1983, 255–89.

International Differences in Wage Behavior: Real, Nominal, or Exaggerated?

By George A. Kahn*

The hypothesis of nominal wage stickiness in the United States and real wage stickiness in the large European countries and Japan has almost become a "stylized fact" of comparative international macroeconomics. As such, the hypothesis has generated a growing number of articles on the implications of and explanations for differing wage behavior across countries. The obvious importance of the hypothesis is its challenge to universal theories of wage determination and its alleged predictions about economic performance. Among the puzzles said to be solved by the nominal vs. real wage adjustment dichotomy are the rapid growth of European unemployment during the 1970's, the reluctance of European and Japanese policymakers to engage in expansionary policy especially after 1973, the greater severity of adverse supply shocks on the European and Japanese economies relative to the U.S. economy, and the worsening U.S. current account balance after 1981. All of these results follow directly from an aggregate supply and demand analysis where nominal wage rigidity induces an upward-sloping supply curve and real wage rigidity induces a vertical supply curve.

A proximate cause of international differences in supply elasticities may be international differences in labor market institutions. In the United States, long-term staggered wage contracts supposedly introduce substantial inertia into nominal wage growth and allow policymakers to alter real wage growth through their control over inflation. Changes in the rate of real wage growth then potentially induce changes in production and employment. In the European countries and Japan, where wage negotiations for most industries occur simultaneously and frequently, the nominal wage is flexible. Any change in inflation passes quickly into a faster rate of increase of nominal wages. In the presence of adverse supply shocks, the customary real wage allegedly established by European labor markets magnifies the real and inflationary consequences of reduced productivity growth.

That institutions may influence wage behavior is not surprising. A more interesting question, though, is what causes workers and firms in different countries to develop differing institutional arrangements to attain presumably similar objectives. Various explanations have been proposed to explain the evolution of institutions causing real wage rigidity in Europe (and perhaps Japan) and institutions causing nominal wage rigidity in the United States. These explanations include the relative importance of supply shocks as opposed to demand shocks, the degree to which an economy is small and open, and the cultural and sociological factors that influence the costs and benefits of wage negotiations and contracts.

This paper takes another look at evidence on wage behavior in Europe, Japan, and the United States in order to assess the accuracy of the influential "stylized fact" of differing wage structures. Section I briefly reviews key institutional characteristics of labor markets in six large industrialized countries. Section II presents a model of wage growth determination sufficiently general to test the influence of these diverse institutions on wage behavior in the manufacturing sector. Estimates of the model, reported in Section III, show that the United States may not be as

*Economic Research Department, Federal Reserve Bank of Kansas City, Kansas City, MO 64198. This paper draws partly on ch. 3 of my unpublished Ph.D. dissertation (Northwestern, 1983). I am indebted to Kermit Daniel for invaluable research assistance. Discussions with Robert Gordon and Jeffrey Sachs are gratefully acknowledged. The views expressed here are my own and do not necessarily reflect the views of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

unique in its wage behavior as previously thought. Variations of the model indicate that a key parameter is unstable with respect to small changes of specification. Depending on the value of that parameter, nominal wage stickiness may or may not be a characteristic of U.S. wage behavior.

I. Wage Contract Characteristics

Before developing a model of wage determination, it is useful and important to examine some of the institutions that determine contract characteristics. Many of the international differences in long-term contracts have been catalogued by Anne Braun (1976), Jeffrey Sachs (1979), and Robert Flanagan et al. (1983). At one extreme, contracts in the United States tend to be staggered by industry and generally last for three years. Furthermore, with respect to indexation, "[t]he elasticity of wage increases to changes in the price level is typically quite low (0.5 or less) in the clauses of major agreements" (Sachs, 1979, p. 319).

At the other extreme is Italy where a system of fully indexed wages, the *scala mobile*, has insulated workers from unanticipated inflation since the institution's inception in 1973. The combination of inflation-neutral real wage growth and accommodative monetary policy has caused a permanent acceleration of inflation in response to real shocks, particularly the oil shocks of 1973–74 and 1979–80. Also, a system of synchronized, plant-based bargaining occurring in three-year cycles has made Italy "the only Western country where wages kept rising [at least after 1975] more than consumer prices even in the years when inflation topped twenty percent and GNP dropped" (Leo Wollemborg, 1982, p. 19). It is interesting to note, however, that since 1980, increasing unemployment, accelerating inflation, and other economic ills have begun to shift the balance between management and labor back toward management, which has always opposed pervasive indexation (Wollemborg).

In between the extremes of institutions contributing to nominal wage stickiness and those contributing to nominal wage flexibility are the types of contracts found in France, Germany, and the United Kingdom. Typically, these contracts specify fixed nominal wages for either unspecified or relatively short periods—usually a year—with varying degrees of effective indexation. In Germany, for example, "the Monetary Law of 1948 as interpreted since 1961" has prohibited "automatic adjustment clauses but not provisions for automatic renegotiation of contracts when the cost of living exceeds a certain level" (Braun, p. 264).

In contrast to European and U.S. wage contracts which are fixed with respect to any contingency other than, perhaps, inflation, Japanese labor market institutions allow wages potentially to respond to a wide range of contingencies. In particular, a semiannual bonus system provides management flexibility in altering wages to match current profitability. Furthermore, the simultaneous renegotiation of wage contracts during the "spring wage offensive" allows annual opportunities for adjusting economywide wage rates. Finally, "a seniority wage, or *nenkō*, system" determines wage payment solely on the basis of "length of service with a company, regardless of such qualifications as skill, past experience, position, and degree of responsibility" (Walter Galenson and Konosuke Odaka, 1976, pp. 607–09). The seniority wage thus encourages lifetime employment and reduces firm-specific training costs. An examination of these features has led Robert Gordon to conclude that "[m]any of the labour-market arrangements selected by the economic theorists to achieve macroeconomic efficiency and high productivity appear to correspond rather closely to well known features of the Japanese labour market" (1982, p. 34).

While institutional descriptions highlight potential areas of cross-country differences in wage behavior, they may not tell the complete story. Wage data for the manufacturing sector include the compensation of workers in a significant number of nonunion firms. In these nonunion firms, unobservable implicit contracts may govern wage behavior and, in the aggregate, may dampen the importance of explicit wage contracts. Furthermore, as has recently become apparent in the United States, explicit contracts are subject to premature renegotiation, especially in times of economic stress. Thus, while observed contracts establish important parameters to be estimated, the wage equation formulated in the next section does not place restrictions on those parameters. Instead, parameter estimates are used to determine the significance of cross-country differences in wage behavior.

II. A Model of Wage Growth

The model to be estimated consists of a partial adjustment equation and a stochastic target wage equation. Observed wage changes are assumed to adjust slowly to changes in target wage growth. Let w_t^* represent the target growth rate of wages at time t. Then observed wage growth, w_t, adjusts as follows:

$$(1) \qquad w_t - w_{t-1} = \lambda\left(w_t^* - w_{t-1}\right).$$

If w_t^* is taken to be the rate of wage growth specified in newly negotiated contracts, then λ can be interpreted as an indicator of contract length.[1] If λ equals one, then w_t^* equals w_t in each period and nominally fixed long-term contracts do not exist. (Equivalently, everyone renegotiates wage growth each period.) As λ becomes closer to zero, however, nominally fixed long-term contracts become more prevalent. The closer λ is to zero, the fewer the number of agents renegotiating contracts each period and the greater the degree of nominal wage stickiness.

The specification of w_t^* assumes agents negotiate a growth rate for expected real wages as follows:

$$(2) \qquad w_t^* - p_t^e = \alpha_0 + \alpha_1 D + \theta\left(p_t - p_t^e\right) + \phi Q_t^e + \beta z_t + \varepsilon_t,$$

where α_0 represents "equilibrium" expected real wage growth, $\alpha_1 D$ represents a permanent shift in equilibrium real wage growth resulting from a shift in productivity behavior, p_t^e represents expected inflation, p_t represents actual inflation, Q_t^e represents expected labor market tightness, z_t represents a vector of supply shocks, and ε_t represents a behavioral error with zero mean and finite variance.

The parameter θ varies from zero to one and indicates the degree of indexation. In observed contracts, indexation clauses insulate agents from price surprises through a variety of techniques. Cost-of-living adjustments, for example, may occur semiannually and be based on inflation rates lagged a quarter. Or, as in Japan, they may take the form of semiannual bonuses, the size of which depend to a greater extent on current market conditions. Thus it becomes important to incorporate into contract equations unanticipated price changes over a period long enough to include lagged cost of living adjustments. Furthermore, since such adjustments usually occur less frequently than every quarter, a time interval longer than a quarter—a year in the present case—seems necessary for the definition of relevant inflation surprises. Indexation to contingencies other than inflation is not permitted in (2) although (2) could easily be generalized to insulate agents from other unexpected outcomes. For Japan, this type of generalization may be important.

Substituting (2) into (1) and rearranging terms gives the equation to be estimated:

$$(3) \qquad w_t = \lambda\alpha_0 + \lambda\alpha_1 D + \lambda p_t^e + \lambda\theta\left(p_t - p_t^e\right) + \lambda\phi Q_t^e + \lambda\beta z_t + (1-\lambda)w_{t-1} + \lambda\varepsilon_t.$$

Two cases emerge under differing assumptions about λ. First, if λ is less than one, then nominal wages are sticky. Both anticipated and unanticipated increases in inflation cause real wage growth to decline. Second, if λ equals one, then real wage growth is neutral with respect to anticipated changes in inflation, as well as with respect to a portion, θ, of unanticipated inflation. In this case, (3) can be reinterpreted as a real wage growth equation. Short-run real wage stickiness will then be indicated if real wages do not respond to excess demand in the labor market or if real wages do not respond to supply

[1] Sachs uses this approach. William Branson and Julio Rotemberg (1980) adopt partial adjustment equations for both nominal and real wages.

shocks. In other words, with λ equal to one, and either ϕ or β or both equal to zero, real wages are sticky in the short run.

Long-run real wage stickiness will be indicated if λ equals one and wages do not respond to changes in "equilibrium" real wage growth, that is, if α_1 equals zero. This result follows because α_1 is the coefficient on D, a dummy variable designed to capture the effect of a long-run decline in productivity growth. D equals zero in the "fast" productivity growth regimes of the 1960's and one in the "slow" productivity growth regimes of the mid-1970's to early 1980's.[2] Thus, a negative value of α_1 indicates that equilibrium real wage growth declined in response to declining productivity growth.

TABLE 1—NOMINAL WAGE GROWTH EQUATION: 1963–81[a]

Parameter	Japan	United Kingdom	United States
α_0	10.025[b]	3.523[b]	4.302[c]
	(1.923)	(0.704)	(2.263)
α_1	−7.141[b]	0.096	−6.456
	(2.517)	(1.183)	(3.814)
λ	0.691[b]	1.296[b]	0.995[b]
	(0.197)	(0.271)	(0.359)
θ	1.676[b]	1.101[b]	4.180
	(0.460)	(0.294)	(3.677)
ϕ	1.185[b]	0.131	0.241
	(0.480)	(0.652)	(1.028)
Mean Square Error	6.082	9.675	41.879

Sources: Hourly compensation in manufacturing from original sources, compiled by the Bureau of Labor Statistics. All other series are from International Monetary Fund, *International Financial Statistics.*

[a] Standard errors are shown in parentheses.

[b] Significant at 5 percent.

[c] Significant at 10 percent.

III. Estimation

Estimation of equation (3) is carried out jointly with the estimation of an inflation equation. Fitted values from the inflation equation proxy for expected inflation in the wage growth equation. Inflation is determined as a linear function of variables in an information set including lagged inflation, output, money growth, and supply shocks.[3] Expected labor market tightness, Q_t^e, is proxied by actual deviations from trend of real *GNP* or *GDP*.[4] The change in the unit value of imports is the principal supply variable and its expectation is also proxied by realized values. Because forecast errors associated with Q_t^e and expected import price inflation may be correlated with each other as well as other right-hand side variables, estimation is by nonlinear three-stage least squares. Instruments include all predetermined variables and lagged money growth. In effect, rational expectations are assumed.

Table 1 reports results for three of six countries sampled, using annual data from 1963 to 1981. The discussion that follows includes results from all six countries—France, Germany, Italy, Japan, the United Kingdom, and the United States. The most surprising result is that the estimate of λ is insignificantly different from one in all countries, including the United States. As a result, equation (3) can be interpreted universally as explaining real wage growth. Inflation indexation is insignificant in all countries except Japan and the United Kingdom. Surprisingly, it is insignificant in Italy where institutional data say it should be important. Only Japan has a real wage that is responsive to excess demand in the labor market, as indicated by a significant ϕ coefficient. In no country does import price inflation signifi-

[2] Dates for the transition from "fast" to "slow" productivity growth are determined by a series of likelihood ratio tests—one for each possible sample split—on the stability of trend growth in productivity. Dates of transition range from 1970 for Japan to 1976 for France.

[3] Two lagged values of inflation and money growth are entered initially. Supply shocks include the change in the real unit value of imports, dummy variables for periods of governmental intervention in the wage setting process, and an autonomous wage push variable for France in 1968. These variables were generally insignificant in the wage equation.

[4] This variable is detrended using quarterly data and a segmented regression model which joins two quadratic trends at a point that minimizes the sum of square residuals (see my 1983b paper). Quarterly residuals from the regression are averaged and used as data. Proper detrending is important because output proxies for labor market tightness through an Okun's law relationship which holds productivity growth constant. Unemployment is not used as the labor market variable because of its unreliability for some countries and in order to avoid having to estimate the natural rate.

cantly affect real wage behavior (coefficients on import price inflation are not reported). Finally, real wage growth in all countries except the United Kingdom declined in response to the productivity growth slowdown. This decline was significant in France, Japan, and the United States. Thus, real wage stickiness, especially that of the cyclical variety, seems to be more or less a characteristic of wage behavior in all countries except Japan.

How can the surprising result that λ equals one in the United States be brought into line with previous research indicating a slow U.S. nominal wage adjustment process? Two modifications to the basic model lead to results more consistent with previous research. First, eliminating the insignificant (and inappropriate) unanticipated inflation term from the U.S. wage equation and reestimating it with two stage least squares changes the λ coefficient to 0.89—a level that is lower, but still not significantly different from one. Second, substituting lagged inflation for lagged wage growth on the right-hand side of (3) to bring the wage equation closer in line with Sachs' (1983) model, reduces the λ coefficient to 0.56—a lower estimate that is insignificantly different from zero (*and* one). Substituting unemployment for output does not change the qualitative results.

The point is not that λ is one or that λ is less than one, but rather that estimates of λ are sensitive to small changes of specification. In fact, with sufficient experimentation, it is not too difficult to generate estimates of λ that are significantly less than one (see, for example, Sachs, 1979, 1983, and my 1983a article). However, these estimates are often unstable over time and depend to a large extent on post-1972 data for the conclusion that nominal wages are sticky in the United States (see my 1983b paper). Additional research is needed to determine the appropriate wage adjustment process and its implication for macroeconomic performance. Better structural models may lead to more reliable estimates of inherently noisy wage equations.

REFERENCES

Branson, William and Rotemberg, Julio, "International Adjustment With Wage Rigidity," *European Economic Review*, May 1980, *13*, 309–32.

Braun, Anne, "Indexation of Wages and Salaries in Developed Economies," *IMF Staff Papers*, March 1976, *23*, 226–71.

Flanagan, Robert, Soskice, David and Ulman, Lloyd, *Unionism, Economic Stabilization, and Incomes Polices: European Experience*, Washington: The Brookings Institution, 1983.

Galenson, Walter (with Odaka, Konosuke), "The Japanese Labor Market," in Hugh Patrick and Henry Rosovsky, eds., *Asia's New Giant: How the Japanese Economy Works*, Washington: The Brookings Institution, 1976, 588–671.

Gordon, Robert, "Why U.S. Wage and Employment Behavior Differs From That in Britain and Japan," *Economic Journal*, March 1982, *92*, 13–44.

Kahn, George, (1983a) "Wage Behavior in the United States: 1907–1980," *Economic Review, Federal Reserve Bank of Kansas City*, April 1983, *68*, 16–26.

______, (1983b) "Nominal and Real Wage Stickiness in Six Large OECD Countries," Federal Reserve Bank of Kansas City Working Paper, August 1983.

Sachs, Jeffrey, "Wages, Profits, and Macroeconomic Adjustment: A Comparative Study, *Brookings Papers on Economic Activity*, 2:1979, 269–332.

______, "A Report on Real Wages and Unemployment in the OECD," *Brookings Papers on Economic Activity*, 1:1983, 255–89.

Wollemborg, Leo, "Italy's 1975 Wage Accord Breaks Down," *Wall Street Journal*, July 7, 1982.

Cross-Country and Cross-Temporal Differences in Inflation Responsiveness

By CHARLES L. SCHULTZE*

This paper compares inflation behavior across widely separated time periods (extending back into the nineteenth century) and among a number of countries in an effort to shed some light on two questions: (*i*) To what extent do the similarities and differences in inflation behavior conform to or depart from the predictions of competing theoretical explanations with respect to such behavior? (*ii*) Is the short- and intermediate-run behavior of prices and wages in the United States more inflexible than in other advanced economies? The countries examined were the United States, Germany, the United Kingdom, Italy, and Sweden. Some of the major results and their implications are first summarized below. The evidence is then presented.

I. Major Conclusions:

In the United States the short-run responsiveness of inflation in the *GNP* deflator to demand shocks was already low at the turn of the century and did not change significantly between the pre- and the postwar periods.[1] This provides at least some evidence for arguing: 1) the postwar stickiness of U.S. inflation is not due solely or perhaps principally to the postwar introduction of three-year staggered wage contracts in the unionized sector of the economy; 2) the reason for the postwar U.S. stickness cannot be attributed to an expectational pattern set in motion by the postwar countercyclical monetary regime; the same stickiness was evident during the prewar gold standard regime.

Inflation in the United States does appear to respond more sluggishly to demand shocks than is the case in the other countries studied, except the United Kingdom. But it has been so for a long time, dating back to the turn of the century. Thus, the current stickiness in the United States relative to other countries does not appear to stem principally from recent institutional changes or international differences in policy regimes.

There is some evidence (from other authors) that several inflation response mechanisms contribute to these results: nominal wages in Europe move more quickly in response to demand pressures; they also move much more quickly in response to changes in consumer prices. The second of these characteristics also helps produce the result that real wages in Europe tend to be slower in making the necessary adjustments to supply shocks than they are in the United States.

II. Variance Analysis

Table 1 summarizes measures of the annual variance of the inflation rate and of "adjusted" nominal and real *GNP* growth in the five countries during two peacetime periods: 1891–1914, the longest period before World War I for which relevant data for all five of the countries were available; and 1953–68, the first half of the period after the Korean War (the second half of the period is examined later). Adjusted nominal and real *GNP* growth are obtained by subtracting the "trend" or long-term growth rate of real *GNP* from the annual growth rate of both nominal and real *GNP*.[2] Thus changes in the

*The Brookings Institution, 1775 Massachusetts Avenue, NW, Washington, D.C. 20036.

[1] In an earlier paper (1981), I found that the flexibility of inflation in nonfarm finished goods prices—as measured by the private nonfarm deflator and the nonfood *CPI*—was approximately unchanged between pre- and postwar periods; the responsiveness of wage inflation to changes in nominal *GNP* declined moderately; and the responsiveness of wholesale price inflation fell sharply.

[2] For prewar years the trend was estimated from a long-term growth rate of real *GNP*, anchored at several points in each country to catch breaks in the trend. In

TABLE 1—VARIANCE AND COVARIANCE ANALYSIS

	Variance		Cov($p\hat{y}$)
	$\hat{y}$	p	Var($\hat{y}$)
1891–1914			
U.S.	44.5	6.8	.22
U.K.	6.0	2.6	.23
Germany	11.0	5.4	.33
Italy	28.5	9.2	.36
Sweden	23.3	7.4	.13
1953–68			
U.S.	6.3	0.9	.21
U.K.	1.9	1.9	.42
Germany	6.5	1.7	.25
Italy	4.9	4.1	.65
Sweden	2.9	3.7	.80

growth trend do not affect the measure of variance. Since nominal *GNP* growth ($\hat{y}$) is the sum of inflation (p) and real *GNP* growth ($\hat{q}$) its variance can be decomposed into two elements—its covariance with inflation and its covariance with real *GNP* growth; that is, $\mathrm{Cov}(p\hat{y})+\mathrm{Cov}(\hat{q}\hat{y})=\mathrm{Var}(\hat{y})$. Alternatively, the simple regression coefficients of p on $\hat{y}$ and of $\hat{q}$ on $\hat{y}$ add to 1.

Strikingly, the annual variance of inflation in the United States in the period immediately following World War II was only one-eighth as great as its variance around the turn of the century. But the much reduced variance of inflation in the later period is not the result of a much smaller inflation responsiveness to cyclical swings in aggregate demand. The covariance ratios indicate that in both periods the covariance of inflation with nominal *GNP* growth was small and very similar. In both periods, only one-fifth of the annual changes in U.S. nominal *GNP* growth were associated with changes in inflation.

By the end of the nineteenth century, the response of inflation to change in nominal *GNP* growth—as measured by the covariance ratios of Table 1—had also become rather sticky in all the European countries. (Covariance ratios in earlier years, 1872–90, for which data are available in Europe but not the United States, had been a good bit higher.)

There was a very wide difference among countries in the magnitude of the variance in inflation. Some of the difference was associated with differences in the magnitude of cyclical swings in *GNP*. While the covariance ratios $\mathrm{Cov}\,p\hat{y}/\mathrm{Var}\,\hat{y}$ were not the same, their differences explained little of the difference in inflation variance among countries.

Relatively simple augmented Phillips curves, with both level and rate of change variables, were also fit by ordinary least squares in the five countries for the years 1891–1914. The equations, very similar to those used by Robert Gordon (1983), took the form:

$$(1)\qquad p=a_0+a_1\hat{y}+a_2GAP(-1)+a_3ri(-1)+a_4p(-1),$$

where p and $\hat{y}$ have already been defined, *GAP* is the percent difference between actual *GNP* and the trend, while *ri* is the annual percent change in relative import prices, (available only for the United States, Germany, and the United Kingdom).[3] The results were used to simulate for each country the short-run response of inflation to two different demand shocks: 1) nominal *GNP* growth was changed by 1 percentage point and kept at the new rate for two consecutive years; and 2) the *GAP* was changed by 1 percentage point and held at the new level for two consecutive years. The inflation response in the second year was then calculated.

Because the trends in real *GNP* used to calculate the prewar *GAP* variable may have been anchored at periods of different degrees of resource utilization, a steadily increasing

the postwar years the U.S. trend growth is taken from the BEA series on potential *GNP* (adjusted by the author to let the benchmark unemployment rate rise gradually to 6 percent by 1980); for European countries the postwar trend growth was based on a nine-year moving average of real *GNP*, interpolated across recession years, and adjusted in beginning and terminal years to exclude both the immediate postwar recoveries and the 1981–82 recession.

[3] In no country was the coefficient on the lagged dependent variable significant, thus it was dropped from the final version and in the simulations reported below.

TABLE 2—CHANGE IN INFLATION IN SECOND YEAR

	Simulated from $\Delta \hat{y}$	Simulated from ΔGAP
U.S.	.45	.34
U.K.	.41–.48	.27–.39
Germany	.58–.63	.55–.67
Italy	.62–.74	.55–.96
Sweden	.66–.83	.74–1.16

or decreasing bias could have been introduced into the measure of the *GAP* variable. An attempt was made to correct for this by introducing time trends into equation (1), coinciding with the periods across which the trends in real *GNP* had been calculated. In those cases where this alternative introduced an important difference into the other coefficients, *both* sets of coefficients were used in the simulation. The results are shown in Table 2.

Two major conclusions are suggested by the table. First, in the years around the turn of the century the U.S. inflation response was somewhat lower than in the other countries, outside of the United Kingdom. Germany, Sweden, and Italy at the turn of the century had much larger agricultural sectors, related to the size of their economies, than the United States and even more so compared to the United Kingdom. This may explain some of the difference in inflation responsiveness. Second, in all countries, except possibly Sweden and Italy, inflation had already become relatively sticky. The output loss required to produce a drop in inflation had become very substantial. Moreover, as discussed later, the use of $\hat{y}$ as the rate of change variable may have introduced an upward bias into the simulated inflation responses. Supply shocks occurring in economies whose aggregate demand curve had a price elasticity between zero and unity would tend to generate a spuriously high coefficient of p on $\hat{y}$. Substitution of $\hat{q}$ for $\hat{y}$ as the rate of change variable did, in fact, eliminate the coefficient on that variable for the European countries and lowered the simulated second-year inflation responses for all countries, while keeping the relative rankings roughly the same (except in Italy for which the use of $\hat{q}$ produced zero or negative coefficients on both the level and rate of change variables).

Finally, in cross-country comparisons, none of the measures of prewar inflation responsiveness shown in Tables 1 and 2 exhibited any significant association with the size of the variance of nominal *GNP*. In general, a positive relationship would have been expected on the terms of the rational expectations *cum* misperceptions model (Robert Lucas, 1973).[4]

III. Early Postwar Experience (1953–68) Europe vs. the United States

The wage explosion in Europe at the end of the 1960's and the oil shocks of the 1970's create difficult problems of identification for inflation equations. Without a complete structural model, it is impossible to know the extent to which changes in nominal *GNP* are exogenous or due to the combination of supply shocks and less than unitary price elasticity of the aggregate demand curve. Initially, therefore, an attempt was made to finesse this question of identification by confining the analysis to the earlier postwar years, 1953–68, when supply shocks were less important.

In this period, by a number of measures, the response of inflation to demand shocks in the European countries (outside of Germany) seems to have been much larger than in the prewar years and much larger than it concurrently was in the United States. The European covariance ratios (Cov$p\hat{y}$/Var$\hat{y}$) were substantially greater than their own values in the prewar period and also much larger than in the United States. When equation (1) was fit to the data for 1953–68 and the two demand shock scenarios simulated, the second-year inflation responses in Europe were much larger than in the prewar period and very much larger than in the United States (again, except for Germany whose response was smaller than the United States).

A number of other studies, investigating various aspects of comparative behavior,

[4] See Richard Froyen and Roger Waud (1983) for a similar finding based on intertemporal comparison of postwar U.S. data.

seem to confirm that in the postwar period U.S. inflation has been stickier than in other industrial countries. It has been unambiguously demonstrated in a number of studies that U.S. wages respond to prior price changes (or changes in price expectations proxies) more sluggishly than is the case for other advanced countries. Most investigators also find that nominal wages in other countries are more sensitive to demand shocks than in the United States. Gordon (1982, 1983) finds this for the United Kingdom and Japan. Jeffrey Sachs (1979) found it for a number of European countries, Japan and Canada. In a later study, Sachs (1983) continued to find a larger nominal wage response for Germany, Japan, and Canada (relative to the United States), but a similar one for France and a lower one for the United Kingdom.

Fitting reduced-form equations very similar to those of equation (1), Gordon (1983) found the postwar inflation response of the United Kingdom and Japan to be substantially greater than in the United States. In a more comprehensive study of 24 countries, for the years 1952–81, David Coe and Gerald Holtham (1983) also fit equations very much like those of equation (1) (except, unfortunately, that the import price variable was left out). A two-year demand shock simulation, using their equations, gave the same basic results as reported above for the shorter period 1953–68: after two years of a 1 percent change in nominal *GNP* growth, the U.S. inflation response is a little larger than Germany, but substantially smaller than the United Kingdom, Italy, and Sweden. In fact, of the 24 countries, only Germany and Switzerland have smaller responses to a change in nominal *GNP* growth than the United States.

Unfortunately, the measures of inflation responsiveness for the earlier postwar period (1953–68) reported above (and, a fortiori, analyses covering later years, featured by large supply shocks) are still plagued by identification problems. As Table 1 reveals, the variance of nominal *GNP* in all of the European countries except Germany was very small in the first half of the post-Korean period. Compared to the 1891–1914 period, for example, the postwar variance of nominal *GNP* was one-third as large in the United Kingdom, one-sixth as large in Italy, and one-eighth as large in Sweden (see Table 1). And nominal *GNP* in all three countries exhibited much less variance than in the United States. The inflation component of that small nominal *GNP* variance ($\mathrm{Cov}\,p\hat{y}/\mathrm{Var}\,\hat{y}$) was quite large, (0.4 in the United Kingdom, 0.7 in Italy, and 0.8 in Sweden). But this large inflation covariance does *not* necessarily imply that the short-run aggregate supply curve in these countries had become very steeply sloped. Rather, when the variance of nominal (and real) *GNP* is very small even small supply shocks can create difficult identification problems. If monetary policy tends partially to accommodate exogenous price increases, a series of small supply shocks, not easily captured in the macroeconomic time-series, will generate a high ($\mathrm{Cov}\,p\hat{y}/\mathrm{Var}\,\hat{y}$) ratio that reflects shifts in the aggregate supply curve along a relatively stable aggregate demand curve rather than tracing out the aggregate supply curve itself.

The variance and covariance relationships support this interpretation. If monetary accommodation is only partial, small positive supply shocks will be associated with small *declines* in adjusted real *GNP* growth, generating a large negative covariance ratio of inflation and real *GNP* ($\mathrm{Cov}\,p\hat{q}/\mathrm{Var}\,\hat{q}$). And that is precisely what the 1953–68 data show for the three European countries. ($\mathrm{Cov}\,p\hat{q}/\mathrm{Var}\,\hat{q}$) is -0.8, -0.5, and -0.4 for Sweden, the United Kingdom, and Italy, respectively, vs. a $+0.10$ for the United States over the same period.

In West Germany the problem of interpretation is quite different. From 1950 to 1960 —before the Berlin Wall was built—the West German labor supply curve was extremely elastic fed by East Germans crossing the frontier. As a consequence, despite sizeable variance in $\hat{y}$ and $\hat{q}$—compared to the other three countries—the response of inflation was very low. But this reflected a *long-run* labor supply curve that was temporarily very flat. The apparently smaller German inflation response in the 1953–68 period may not at all represent movement along a short-run disequilibrium aggregate supply curve, and the coefficient on $\hat{y}$ may *understate* the German inflation responsiveness.

Additional problems arise in interpreting the coefficient on the lagged dependent variables, when the analysis is extended into the 1970's. The coefficient on this variable (in various versions of equation (1)) was insignificant in the prewar period in almost all cases, while generally taking on significant positive values in the early postwar period (except for Sweden), ranging from 0.26 in Germany to 0.72 in the United States. One interpretation of this phenomenon, which I favor, is that norm rates of inflation (to use an Okun-Perry concept) or long-term inflation expectations (to use a more conventional concept) do not adjust to short-lived, high-variance cyclical changes in demand and inflation, such as characterized the prewar period, but do move in response to long sustained movements in the demand and inflation. (During the period from the Civil War to World War I the mean length of expansions was 24 months and of contractions 23 months; from 1945 to 1980 the corresponding numbers were 49 and 10.) The initial upward adjustment of the overall price level accompanying large supply shocks is likely to get at least partially incorporated into the inflation norm (or inflation expectations). Unless allowance is made for this fact —by more than the inclusion of relative import prices, energy prices, or similar variables—the coefficient on the lagged dependent variable may pick up a higher value than is likely to be valid for simulating the impact of less dramatic demand or supply adjustments.

To deal with these problems, equation (1) was modified.

Two dummies were entered: $D1$ for European countries, to reflect the late 1960's wage explosion, taking on values of 1 after 1967; $D2$ taking on values of 1 after 1973 to capture possible nonlinear expectational (or norm-changing) effects of the oil-supply shocks. The coefficients on $D2$ were not significant for Germany, and the United Kingdom. The U.K. equation was estimated only through 1978, to avoid picking up the effect of the VAT increase in 1979. The German equation was fit from 1960 through 1980 to avoid the period of large East German migration.

TABLE 3—MAGNITUDE OF SECOND-YEAR INFLATION RESPONSE TO DEMAND SHOCKS

	1% Shock to $\hat{y}$		1% Shock to *GAP*	
	With $\hat{y}$	With $\hat{q}$	With $\hat{y}$	With $\hat{q}$
U.S.	0.56	0.45	0.55	0.41
U.K.	0.99	0.26	2.36	0.26
Germany	0.71	0.66	0.86	0.74
Italy	1.02	0.91	2.59	1.60
Sweden	0.96	0.84	1.94	0.84

Two versions of equation (2) were then estimated for 1953–80, one with $\hat{y}$ as the rate of change variable, another with $\hat{q}$ as that variable. While the coefficients on $\hat{y}$ are likely to have been biased up, at least in the United Kingdom, Italy, and Sweden, the coefficients on $\hat{q}$ may be biased down, to the extent that supply shocks impinged on only partially accommodative aggregate demand policy. In the case of Sweden and the United Kingdom, the coefficients on $\hat{q}$ were in fact negative, and the equations were refit and simulated without the $\hat{q}$ variable.

The results of carrying out the two demand shock simulations for each version are shown in Table 3. The estimates should bracket the correct ones.

In the United States, it makes little difference which version is used—the inflation responses are approximately the same and close to the prewar responses reported in Table 2. (They are also similar to the results obtained by fitting the $\hat{y}$ version for the 1953–68 period.) The results, on either version, broadly confirm the earlier findings that the short-run flexibility of inflation is less than in most European countries. But the degree to which the inflation response in Europe exceeds that of the United States is reduced, in some cases very much so, for three of the four European countries when $\hat{q}$ is substituted for $\hat{y}$. Moreover the positions of the United Kingdom and Germany are reversed. The United Kingdom is seen to have substantially stickier inflation than the United States, while the German inflation response is now larger when the 1953–1959 years are eliminated, as was suggested by the earlier discussion.

The revised position of Germany in the ranking is also broadly consistent with the more casual observation that Germany appears to have been able during the 1970's to contain the spread of supply shock inflation with demand management tools at lower cost than the United States.

Finally, when $\hat{y}$ is used as the rate of change variable, there is a positive and significant association across countries between the variance of adjusted nominal *GNP* and the size of the inflation response under either simulation. But, as expected, when $\hat{q}$ is substituted for the rate of change variable, the association, while still positive, becomes less than significant in the two simulations. On balance, taking both prewar and postwar data into account, the evidence (from this very small sample) on the relationship postulated by Lucas is at best mixed.[5]

While differences in the nature of labor contracts in the unionized sector of the five economies may explain some of the international differences in inflation behavior, the evidence suggests that this is far from the sole or perhaps primary cause. The fact that the United States is a much less "open" economy than the others may be a reason—although the United Kingdom must then be treated as a special case. Additional research directed towards institutional sources of differential wage and price behavior seems in order.

[5] Coe and Holtham find a significant positive association across 24 countries between the coefficient on $\hat{y}$ in their equations and the variance of $\hat{y}$. However there is, as explained above, a serious identification problem with the formulation especially in the postwar period. Coe and Holtham's use of two-stage least squares may not, in my judgment, have avoided this problem.

REFERENCES

Coe, David T. and Holtham, Gerald, "Output Responsiveness and Inflation: An Aggregate Study," Working Paper No. 6, Economics and Statistics Department, OECD, April 1983.

Froyen, Richard T. and Waud, Roger N., "Demand Variability, Supply Shocks and the Output-Inflation Tradeoff," Working Paper No. 1081, National Bureau of Economic Research, February 1983.

Gordon, Robert J., "Why U.S. Wage and Employment Behavior Differ From That in Britain and Japan," *Economic Journal*, March 1982, *92*, 13–44.

______, "A Century of Evidence on Wage and Price Stickiness in the United States, the United Kingdom, and Japan," in *Macroeconomics, Prices and Quantities*, Washington: The Brookings Institution, 1983, 85–133.

Lucas, Robert E. Jr., "Some International Evidence on Output-Inflation Tradeoffs," *American Economic Review*, June 1973, *63*, 326–34.

Sachs, Jeffrey D., "Wages, Profits and Macroeconomic Adjustment: A Comparative Study," *Brookings Papers on Economic Activity*, 2:1979, 269–319.

______, "Real Wages and Unemployment in OECD Countries," *Brookings Papers on Economic Activity*, 1:1983, 255–89.

Schultze, Charles L., "Some Macro Foundations for Micro Theory," *Brookings Papers on Economic Activity*, 2:1981, 521–76.

The Public Capital Stock: Needs, Trends, and Performance

By CHARLES R. HULTEN AND GEORGE E. PETERSON*

The condition of America's roads, bridges, mass transit systems, and water and sewer facilities has become an issue of increasing controversy. One view holds that deterioration of this "public infrastructure capital" has reached alarming proportions, and that a significant fraction of future national savings will be needed to reverse the damage of past neglect. Another view argues that there has been no sudden acceleration of decay in capital stock performance, that better facility management can often substitute for capital investment, and that attempts to greatly increase the size and condition of the public capital stock would place unsustainable pressure on financial markets, state and local government budgets, and the federal system of grants-in-aid.

How extensive *is* the capital infrastructure problem? How did it arise, given the many federal government subsidies available for state-local capital expenditure? How much of the national capacity for capital formation should be allocated to solving the problem? If there is a divergence between actual and desired infrastructure capital, can the gap be closed without disrupting the growth of other sectors and governmental objectives? These questions are, unfortunately, much easier to pose than to answer. There is only limited information about the size and condition of the stock of public infrastructure capital, and much of the available information has been assembled from the perspective of engineering or planning "needs standards." Needs assessments typically indicate the fraction of the infrastructure stock that falls below a specified (and somewhat arbitrary) level of performance, or indicate how much spending would be required to bring the existing stock up to the specified standard. While sometimes appropriate as a management tool, infrastructure needs estimates are formulated in isolation from other public and private capital needs and are thus ill-suited to the policy debate over national priorities.

Despite significant data problems, some general conclusions about the infrastructure dilemma are possible. As a backdrop for these conclusions, the first section of this paper presents an overview of the historical trends in public sector capital formation and infrastructure condition. The following section discusses the factors shaping these trends, with particular attention to the supply and demand components of the capital spending decision. We then consider the question of whether existing trends will insure an adequate supply of public capital, and address the issue of whether a reduced quantity and quality of infrastructure capital may be expected to reduce private sector productivity growth.

I. Historical Trends in Capital Formation and Condition

Economic trends in the state-local sector have followed two distinct patterns. As can be seen in Table 1, the 1950's and 1960's were a period of rapid growth in total expenditures. This growth, as measured by the share of *GNP* devoted to state-local spending, came to a halt in the 1970's and has been gradually declining since 1975.

Spending on structures and equipment shows a somewhat different pattern. The share of total expenditure devoted to gross sectoral capital formation was relatively stable until the mid-1960's, and then declined dramatically. The decline in *net* investment is

*The Urban Institute, Washington D.C., 20037. We thank Nancy Humphrey, Edward Gramlich, and Michael Boskin for valuable comments on earlier drafts of this paper.

TABLE 1—SELECTED TRENDS IN STATE-LOCAL GOVERNMENT SPENDING[a]

Years	State-Local Gov't Spending as a Share of *GNP*[b]	State-Local Capital Spending as a Share of Total Spending[b]	Average Annual Growth Rate of the Stock of Structures[c]
1958–62	9.3	30.8	2.32[d]
1963–67	10.4	29.2	3.15
1968–72	12.3	23.8	2.22
1973–77	13.4	18.6	1.01
1978–82	12.8	15.7	0.26

[a]Shown in percent.

[b]*Source*: The U.S. *National Income and Product Accounts*.

[c]Source: Unpublished estimates underlying Hulten (1984).

[d]1959–62

even more severe. When depreciation is subtracted from gross investment in structures (which accounts for 95 percent of the stock of fixed capital), the resulting estimate of net investment is found to have increased significantly from 1958 to 1968, then to have fallen off precipitously. Under the assumption that public structures depreciate at an average rate of 3 percent a year, net investment in state-local structures has almost reached zero in the 1980's. This is reflected in the last column of Table 1, where it is seen that the stock of structures has virtually ceased to grow.

Data on maintenance spending in the state-local sector is treated as a current expense, and not assigned a separate category in the *National Accounts*. However, analyses of city budgets suggests that maintenance spending also declined in the 1970's. This decline accompanied, and often outpaced, the decline in spending for new capital formation, and is evidently associated with the conscious budget strategy known as "deferred maintenance." The rationale for this strategy is discussed below. We direct attention here to the fact that most types of infrastructure capital have relatively well-defined best-practice maintenance paths, and that excessive deferral of maintenance may cause a serious deterioration in condition. Roads, for example, have cost-effective resurfacing cycles, which depend on the type of surface, initial condition, climate, and snow removal procedures, but which can be defined to a good approximation in any one locality. Bridges must be repainted on a regular basis; mass transit vehicles have normal maintenance cycles; and water and sewer systems have components which deteriorate and need to be replaced. Normal maintenance and repair can be deferred for some time without adversely affecting the condition of the stock, but a point is typically reached beyond which further deferral results in a rapid deterioration of condition.

Many cities during the 1970's deferred maintenance far beyond efficient limits. While comprehensive and comparable data are difficult to obtain, scattered information can be obtained on the condition of most types of infrastructure, and reasonably good information is available for roads. Condition and performance measures for 62 cities were examined by Peterson et al. (1983), who conclude that a deterioration in most types of infrastructure did occur during the 1970's, but that the problem has reached serious proportions in only a limited number of jurisdictions (mainly large cities with severe fiscal crises). In-depth studies of specific cities—the Cleveland study by Nan Humphrey et al., (1982), for example—reinforce this finding.

The overall pattern of capital formation in the state-local government sector is thus one of expansion during the 1950's and 1960's, and decline thereafter. This decline occurred both in the formation of new capital and in the maintenance and condition of existing capital. It is apparent that the sector as a whole chose to use up its capital at a higher rate during the 1970's and early 1980's, creating diminished capital consumption possibilities for the future.[1] The factors underlying this de facto choice are considered in the

[1] It is important to note that a part of today's capital "Needs Gap" comes from newly adopted performance standards that have rendered obsolescent existing public capital. New water treatment standards, for example, account for the bulk of the nation's capital needs for wastewater systems.

following section, and the consequences are discussed in a subsequent section.

II. Explanation of Historical Trends

A. *Theoretical Considerations*

The conventional theory of private sector investment identifies two main components of the demand for capital: the desire to produce a given level of output; and the cost of using capital relative to the cost of other inputs. Formally, the demand for capital by a cost-minimizing firm is a function of the user cost of capital, the price of noncapital inputs, and the desired level of output. Unfortunately, several factors prevent this model from having immediate applicability to the public sector. Governments typically operate under political and institutional constraints that are unknown to private businesses, and cost minimization is but one factor in public sector spending decisions. Furthermore, since public sector output is rarely marketed at marginal cost, data on the cost of goods sold are not available as a guide for measuring output, or as a guide for minimizing costs.

Conventional theory has, nevertheless, proved a useful starting point for modeling public sector production decisions. In a recent paper, Hulten (1984) has extended the Gramlich utility-maximizing model of the public sector to include a "household" production function. This extension allows public sector output to be estimated using econometric techniques, and an a priori estimate of output is not necessary.[2] The demand for capital is specified parametrically, in this model, as a function of the user cost, the price of other inputs, the price of private sector output, the implicit price of public sector output, and the income allocated to the consumption of public goods. This is the public sector analogue of the private sector demand function noted above.

B. *Capital Spending Trends: An Explanation*

This model of public sector investment is useful for interpreting the trends noted in the preceding section: capital stock is demanded in order to produce public sector goods for direct final consumption, and to complement private sector production inputs. Since both types of output are "normal" goods, demand increases with income. In this regard, it is important to note that the 1950's and 1960's were a period of rapid economic growth, and that the derived demand for public sector capital increased accordingly. This general demand effect was augmented by demographic trends which led to an increased demand for specific types of capital: the post-World War II baby boom led to an increased demand for schools; the overall growth in population increased the demand for roads, mass transit, and expanded water and sewer facilities; the shift in population between urban, suburban, and nonurban areas, and the shift between Snow Belt and Sun Belt required capital investment to match the location of the capital stock to the consumer population.

The 1970's altered these trends. Both population and economic growth slowed during the 1970's, and the reduced growth in the demand for public sector services was reflected in part by the "expenditure limitation movement." The overall decline in demand was reinforced by shifts away from the specific public sector goods which had contributed greatly to capital spending: the baby boom worked its way through the school system, reducing the need for school construction, and the Interstate Highway System neared completion.[3] To the extent there was demand for new capital spending in this period, it came largely from the adoption of new standards for public services, which

[2] This model treates state-local governments as though they were households faced with the problem of allocating income between public and private uses. The public sector commodity is not directly observed, but is assumed to be produced with purchases of policemen, teachers, etc. Public output is thus not an observable variable in this model, but can be estimated from data on public inputs and on total expenditures.

[3] The decline in spending for schools and highways explains a large part of the decline in capital expenditure noted in Table 1. However, the share of net and gross capital spending declines even when these two functions are removed from the data.

rendered obsolescent part of the inherited capital stock. Higher standards for wastewater treatment generated replacement demand for treatment plants; the promulgation of width and other standards for bridges triggered rehabilitation investment for bridges.

An even more important shift in spending priorities occurred in the relationship between social welfare and capital components of state-local government budgets. The Great Society programs initiated in the mid- and late 1960's imposed additional spending requirements on state-local budgets, and the 1970's saw a significant increase in the share of the budget devoted to social welfare programs. Given limits, implicit and explicit, on overall spending levels, the crowding out of other state-local programs was inevitable. Capital spending was a natural candidate for the crowding-out process, since, as noted above, it can be deferred for some time before adverse consequences are created (and even longer before these are noticeable). This crowding-out effect was most pronounced in cities experiencing fiscal distress. Studies of infrastructure spending and condition have found that the most acute infrastructure problems are concentrated in such cities.

The capital spending patterns described in the preceding section may thus be explained to a large degree by the shifts in demand for public sector services. It is important to note, however, that at least some of these "demand-side" shifts interact with "supply-side" incentives put in place at the federal level. A full understanding of capital trends at the state-local level thus requires an overview of the specific incentive programs.

C. *The Public Supply Side*

Supply-side incentives influence the demand for investment by reducing the cost of using capital. In the private sector, investment incentives are provided by the federal government primarily through the tax system. This route is obviously not available for state-local government investment since they are tax exempt entities, and federal incentives for capital formation have taken the form of direct grants-in-aid and subsidized municipal bond interest rates.[4]

Federal grants for capital spending provide a potentially powerful incentive to state and local governments to spend on the targeted items. For every dollar of own spending up to a specific limit, a matching grant at rate m provides $m/(1-m)$ dollars of federal funds to the project. The matching rate varies by function: the Highway Trust Fund, established in 1956, provides for a 90 percent matching rate for interstate highways and a 75 percent rate for other roads; the Construction Grants Program for water pollution control established a 30 percent rate in 1956, which was raised to 50 percent in 1966 and to 75 percent in 1972, before being lowered again in recent legislation; for assistance to mass transit, a $66\frac{2}{3}$ percent rate was established in 1964, and raised to 80 percent in 1974; the Bridge Replacement and Rehabilitation Program was established in 1972 with a 75 percent matching rate, which was raised to 80 percent in 1978.

The size of these matching grants suggests a strong incentive to investment in the targeted areas. Typically, when a new matching grant program is introduced, the federal aid level is large relative to state-local spending for the targeted purpose. Thus, the initial effects are highly stimulative. However, almost all federal capital grants are closed ended, and as state-local spending rises over time, the closed-ended ceiling is reached. Grant stimulation for additional investment thus vanishes as federal dollars are substituted at the margin for state-local dollars from own sources. Substitution is especially strong for noncategorical capital aid programs like temporary public works assistance and community development block grants, which became a larger part of the federal capital aid mix in the 1970's. Ironically, real capital aid from the federal government grew

[4] Sale and lease-back arrangements do capture for the public sector some of the federal tax incentives for investment. Under these arrangements, the public sector sells an asset to private owners, then leases back the asset for its own operations at a price which reflects the depreciation and other tax advantages to the private owners.

TABLE 2—THE SOURCES OF FINANCING FOR STATE-LOCAL GOVERNMENT CAPITAL SPENDING[a]

Year	Federal Grants as Percent of Gross Capital Spending[b]	Long-Term Borrowing as Percent of Gross Capital Spending
1960	24.4	37.1[c]
1965	24.9	35.0[c]
1970	24.4	57.4
1975	25.7	49.1
1981	39.9	24.0
1982	37.4	58.4

[a] Shown in percent.

[b] Source: Peterson et al. (1983), except as noted.

[c] Paul Schneiderman (1975). The 1960 and 1965 estimates are not directly comparable to the later estimates.

throughout most of the 1970's (see Table 2) —reaching its highest share of state-local capital spending at the end of the decade and in the early 1980's—during a period when real state-local capital expenditures were falling.

Until 1976, repair and maintenance went almost unnoticed in federal aid design. There were no matching grants for these purposes, and many grants prohibited the use of federal dollars for maintenance. The federal grant structure thus steered spending away from the maintenance of existing facilities and toward new construction. This has been cited as a contributing factor in the deterioration of infrastructure condition.

The tax preference accorded in municipal bonds is the second major way in which state-local capital spending is subsidized by the federal government. Interest received from municipal bonds is excluded from the base of federal level income taxes, and thus permits access to debt finance at interest rates considerably below those prevailing on taxable debt (until the industrial revenue bond boom of recent years, municipal bond rates averaged 25 to 30 percent less than interest rates on comparable private sector debt). Since long-term borrowing is the primary method through which state-local capital formation is financed, the interest exemption for municipal bonds is of major significance to state-local capital spending decisions.

The two funding sources which historically account for more than two-thirds of state-local capital spending are thus heavily subsidized by the federal government. The remaining spending is financed by a combination of user fees and general tax revenue, with the latter allowed as an itemized deduction against the federal personal income tax. The overall pattern of supply-side incentives is relatively favorable to state-local capital formation, and almost certainly has resulted in a larger stock of infrastructure capital than would have resulted without the favorable treatment.[5]

But, do changes in these supply-side incentives explain the pattern of rise and decline noted in Section I? The answer is a qualified no. Toward the end of the period, pressures on the tax-exempt bond market from private uses for industrial investment and home mortgage lending did narrow the historical advantage of tax-exempt borrowing to finance infrastructure investment. But, such nontraditional uses of the municipal bond market are a relatively recent event and cannot explain the decline in capital spending occurring in the early and mid-1970's. It is also true that federal movement from categorical capital aid to general aid made possible greater substitution for state-local own-source funds and blunted the stimulative effect of federal assistance. But, while there has been some erosion in the value of federal incentives, it is impossible to attribute the last decade and a half's decline in capital formation to a loss of federal grant resources. As we have shown in Table 2, federal aid was the best-sustained revenue source during the period of investment decline, and

[5] Income originating in the private business sector is subject to a significantly positive rate of taxation, while the implicit income from state-local capital receives a significant subsidy. According to the forthcoming study of Mervyn King and Don Fullerton, the marginal effective tax rate on corporate income was in the 45–48 percent range in 1960, the 44–47 percent range in 1970, and in the 29–33 percent range under current tax law. Income accruing to owner-occupied residential housing received highly favorable treatment under the personal income tax, but is subject to property taxation.

accounted for a rising share of state-local capital spending during this period.

This is not to say that federal matching grant programs have played no role in shaping capital spending trends. Federal grants have had a profound influence on the composition of state-local capital spending, and the grant-induced expansion and subsequent contraction of the federal highway program is a significant factor explaining the overall spending patterns of Table 1. Furthermore, the recently enacted 5¢ federal gas tax is now promoting a boom in state-local highway and bridge repairs. Finally, the programmatic bias against maintenance was a factor in the declining condition of the infrastructure stock. The structure of federal incentives thus has left its mark on the state-local capital stock and on investment trends, but *changes* in federal incentives do not appear to be primarily responsible for the decline in capital outlays. The explanation for the secular decline in state-local capital spending lies far more on the side of government and taxpayer *demand* for capital outlays, induced by a slower growth in income and by demographic shifts in the age and location of the population.

III. The Adequacy of Public Sector Investment

A. *Capital Needs vs. Capital Wants*

We have seen that capital formation and condition declined during the 1970's, and discussed the factors behind these trends. We have not yet addressed the central issue of the capital needs problem: does the declining spending of the 1970's mean that there was underinvestment in public sector capital, and that the current and prospective future stock is in some sense inadequate? This section is devoted to an attempt to address various aspects of this question.

One possible answer lies in the use of the needs approach mentioned in the introduction. An inventory of infrastructure condition would indicate the portion of the stock falling below prescribed standards. Such an inventory would thus provide an estimate of the shortfall of past investment relative to the prescribed standard, and in some cases the amount of investment required to bring the stock up to standard.[6] Studies based on this approach have estimated that as much as $3 trillion in capital spending would be necessary over the next one or two decades in order to bring the infrastructure stock up to par, although other estimates are of a more modest scope.[7]

The general problem with this approach is that it fails to account for tradeoffs with other spending objectives. Few people maintain their homes or their cars in good-as-new condition because they have other uses for their income. Similarly, the U.S. economy has a variety of capital needs outside the public sector—housing, business plant and equipment—and a limited pool of saving on which to draw. Government also has many demands on its tax revenues and cannot look to the condition of its capital stock in isolation from other spending needs. From an economic standpoint, the correct balance between competing objectives comes when the social rate of return on public sector capital is equated to the social rate of return on other types of capital.

The equality of social rates of return is a standard theoretical proposition in welfare economics, but is a very slippery concept to apply to practical investment decisions. The social rate of return on public sector investment is extremely difficult to estimate, since the services of capital are not usually provided on a marginal cost basis and are thus difficult to value, and because there are usually thought to be externalities associated with the stock of roads, sewers, and mass transit systems. Even if these measurement problems could be overcome, there is the deeper theoretical issue about the appropriate rate of discount.

[6] Many states have launched efforts to carry out formal inventories and needs assessments, and several pieces of federal legislation have been introduced to do the same on a national scale (Joint Economic Committee, 1984).

[7] The $3 trillion estimate is presented in the studies by Pat Choate (1982) and Associated General Contractors (1983). A more moderate estimate of needs is presented in the Congressional Budget Office study (1983).

Despite these difficulties, some general observations are possible. First, the current difficulties in capital stock maintenance to some degree reflect the boom in public capital investment that occurred in the late 1950's and early 1960's. Many of these facilities have reached the end of their planned lives, and now require replacement or more intensive maintenance and repairs. Especially in older cities, governments must wrestle with a capital stock network that is more extensive than they would choose to own if they were deciding to build their capital facilities anew. The period of high economic growth coupled with stimulative grant programs probably promoted *overinvestment*, in the sense that government leaders would have devoted fewer resources to new capital investment if they had foreseen the subsequent economic slowdown or accurately forecast lifetime maintenance costs. The infrastructure problem is misrepresented when interpreted as resulting from long-term, persistent underinvestment in the state and local sector.

How should the backlog of investment "needs" now be reconciled with limited investment resources? In private markets, this conflict is resolved through the price mechanism. Traditionally, service prices have not played much of a role in capital allocation within the public sector, or between the public and private sectors, but some movement of this kind is now taking place. In recent years, a number of local governments have reorganized their sewer and water services to place these on a full incremental-cost fee basis, and to guarantee future fee adjustments to preserve incremental-cost pricing. Other communities have adopted neighborhood voting on special assessments to pay for neighborhood street repairs. These mechanisms not only ensure financing for capital investment and repairs, but act as restraints on taxpayer demand for capital expenditures. The city of Milwaukee, for example, has found that neighborhood households are willing to pay for far fewer street repairs than street department officials believe are "needed" and recommend for the city's capital budget.

There remain important obstacles to having taxpayer or user voting resolve conflicts over the proper levels of capital and maintenance spending. Some states now require more-than-majority voter approval for general obligation bond issues, and these requirements have created a tilt toward underspending. The state of California, for example, requires two-third majorities for approval of general obligation bond issues, virtually eliminating tax-supported debt as a vehicle of capital finance.

Most fundamentally, for public voting on tax or bond proposals to be successful, there must be comprehensible information on the consequences of continuing to defer maintenance or improving capital facility condition. This information for the most part does not exist. The technical experts who conduct needs studies are themselves just beginning to identify the costs or service consequences of alternative maintenance levels. Thus far they have done very little to bring such information before the public in a credible fashion. The lack of reliable information on the consequences today's investment and maintenance decisions will have on tomorrow's capital condition often has reduced the infrastructure debate to conflicting assertions about "need," supported only by historical rules of thumb about the capital standards that should be met.

B. *Infrastructure Condition and Private Sector Productivity*

Attempts to match infrastructure investment decisions with taxpayer or consumer willingness to pay must be examined in light of the argument that the size and condition of the infrastructure stock has significant spillover effects on the productivity of private sector businesses. The spillover from public capital deterioration has been cited in the popular press as a contributing cause of the productivity slowdown in the private business sector. It has also been cited as a factor in the poor economic performance of the Snow Belt region of the United States relative to the newer Sun Belt region.

It is difficult to accept or refute this hypothesis on the basis of existing evidence. While the development of social overhead infrastructure capital is widely accepted as an important factor in economic development, it is much less obvious that variations in the condition of existing infrastructure stock have a major impact on business productivity. The absence of a road network may seriously inhibit economic growth, but this does not necessarily mean that potholes have a similar effect. It is not even clear just how infrastructure capital enters the production function of private business. Attempts to quantify the business costs created by road and bridge deterioration, for example, assign most of the costs to greater commuting time and to wear and tear on commuters' automobiles. Longer or more unpleasant commutes may affect worker productivity, deliveries may be delayed by inadequate roads, and a firm may have to substitute its own capital to compensate for inadequate water or sewer facilities or for wear and tear on its vehicles. But these effects primarily influence the relative attractiveness of alternative business locations. No evidence has been advanced to show that deterioration of public capital has significantly influenced the productivity of the nation's private sector as a whole.

Some indirect evidence suggests that public capital conditions may not significantly influence regional manufacturing growth. A recent study by Hulten and Robert Schwab (1984) examines the causes of the productivity slowdown in the manufacturing industries of the Snow Belt and Sun Belt. The productivity slowdown was found to be common to all regions, implying that the generally poorer infrastructure condition in the Snow Belt reported by Peterson et al. did not translate into poorer productivity growth. The linkage between public capital and private production deserves more attention than it has received, but any hypotheses require careful testing against the evidence, not mere assertions of the importance of deteriorating public sector capital to private output.

REFERENCES

Choate, Pat, "House Wednesday Group Special Report on U.S. Economic Infrastructure," unpublished manuscript, May 1982.

Gramlich, Edward, "Alternative Federal Policies for Stimulating State and Local Expenditures: A Comparison of Their Effects," *National Tax Journal,* June 1968, *21,* 119–29.

Hulten, Charles R., "Productivity Change in State and Local Governments," *The Review of Economics and Statistics,* 1984 forthcoming.

______ **and Schwab, Robert M.,** "Regional Productivity Growth in U.S. Manufacturing: 1951–78," *American Economic Review,* March 1984, *74,* 152–62.

Humphrey, Nan, Miller, Mary John and Godwin, Steve, "Financing Greater Cleveland's Capital Requirements," Washington: Urban Institute, October 1982.

King, Mervyn A. and Fullerton, Don, *The Taxation of Income From Capital: A Comparative Study of the U.S., U.K., Sweden, and West Germany,* Chicago: University of Chicago Press, forthcoming.

Peterson, George E., "Financing the Nation's Infrastructure Requirements," in Royce Hanson, ed., *The Adequacy and Maintenance of Urban Public Facilities,* Washington: National Academy of Sciences, 1984.

______ **et al.,** "Benchmarks of Urban Capital Condition," Urban Institute, February 1983.

Schneidermann, Paul, "State and Local Government Gross Fixed Capital Formation: 1958–73," *Survey of Current Business,* October 1975, *55,* 17–26.

Associated General Contractors of America, *America's Infrastructure: A Plan to Rebuild,* Washington, 1983.

Congressional Budget Office, *Public Works Infrastructure: Policy Considerations for the 1980's,* Washington: USGPO, April 1983.

Joint Economic Committee of the United States Congress, *Financing Infrastructure Renewal and Development: A National Response,* Washington: USGPO, 1984.

U.S. Department of Commerce, Bureau of Economic Analysis, *The National Income and Product Accounts of the United States, 1929–76, Statistical Tables,* Washington: USGPO, 1981.

The Displacement of Local Spending for Pollution Control by Federal Construction Grants

By JAMES JONDROW AND ROBERT A. LEVY*

It is well known that expenditure by the federal government can displace expenditure by private citizens and by state and local governments. Among the possible mechanisms for displacement, one is particularly direct: federal expenditures provide services to recipients that replace, to some extent, what the recipients would have purchased on their own. Food stamps, for example, increase the income of recipients, but not necessarily their food purchases; a substantial part of the aid simply might pay for food the recipient would have bought on his own. The same kind of displacement can occur when the federal government provides grants-in-aid to a state or local government for spending on particular items. If displacement is large, the federal role could be reduced without greatly affecting the total expenditure on those items.

In this paper we measure the extent to which state and local government spending on sewer system construction is displaced by EPA construction grants. The Construction Grants Program provides federal funds to local governments for the construction of sewer lines and sewage treatment plants. (This is a big program, as much as $3–$4 billion in recent years.) The federal role is often justified by referring to the externalities in pollution control; water pollution generated in one place moves downstream and may affect other communities (perhaps in another state). The federal Construction Grants Program was instituted to reduce water pollution by subsidizing the construction of large waste water treatment plants. Though the program provided matching grants for the construction of plants, there were always some funds that were devoted to sewer lines as well.

If pollution is external to the community and state and local governments are not concerned with pollution flowing downstream, then federal expenditures should add to funds spent by local governments and generate little displacement. On the other hand, if federal funds substitute for local funds, perhaps because local governments were already doing something about the problem or because they were able to use the funds in ways to suit their own purposes, then substantial displacement should result.

We consider two kinds of displacement resulting from the grant program: permanent and temporary. Permanent displacement is created by the substitution of grants (after they are authorized by the federal government) for local expenditures that would have served the same purpose.

Temporary displacement is due to postponed spending by state or local governments primarily at the beginning of the program. There are two forms of temporary displacement. First, there are the delays built into the cranking up of any new program and inherent in the process of obtaining grants and meeting procedural requirements for spending them. Second, there is the waiting to see if one's project can get funding. A local politician cannot afford to start a project without federal aid when there is even a suspicion that funding could have been obtained.

I. The Model

We begin with the demand for sewer system structures financed by state and local funds. Note that the demand is in terms of a stock (structures) not a flow (new construction), since it is the entire stock that gen-

*The Public Research Institute of the Center for Naval Analyses, 2000 N. Beauregard Street, Alexandria, VA 22311. Helpful comments have been made by Edward Gramlich and Paul Feldman, and valuable research assistance provided by George Dougherty.

erates services. This stock (S) is created using the perpetual inventory method and includes both sewer lines and treatment plants. The demand for S (denoted by S^*) is directly related to the stock of housing (H):

$$(1) \qquad S^* = \alpha H^{\beta},$$

where β measures the elasticity of S with respect to H.

Equation (1) represents demand in the absence of federal outlays, so that all of S^* is built with the community's own funds. We now generalize the definition of S^* to recognize that demand can be satisfied by structures built with grants as well as those built with the community's own funds. However, we include in S^* only the fraction (γ) of the stock of structures built with grants that serves the same function as the stock built with municipal funds.[1] This fraction can be interpreted as a measure of the "value" the community places on the stock of sewer structures built with federal funds. The fraction will approach unity when the value of the project is high so that federal expenditure displaces most of what the community would have built on its own. Similarly, when γ approaches 0, federal expenditure causes no displacement of local spending. If γ is strictly greater than 0 and less than 1, the remaining fraction ($1-\gamma$) of the stock built with grants represents structures that the community would not have built, perhaps because the facility is of a different type or in a different location from one that the community would have chosen.

We define as the "effective" stock those structures built with the community's own funds (M) plus the share (γ) of the stock of structures built with grants (G) that provides service to the community. The desired effective stock is

$$(2) \qquad S^* = (M+\gamma G)^* = \alpha H^{\beta}.$$

Like all variables measuring desired levels, S^* is unobservable. It is eliminated from the equation to be estimated in a standard way: by incorporating partial adjustment. The assumption of partial adjustment in multiplicative (log-linear) form is

$$(3) \qquad S/S_{-1} = \left(S^*/S_{-1}\right)^{\eta}.$$

Combining (2) and (3) to eliminate S^* yields

$$(4) \quad M+\gamma G = \alpha^{\eta} H^{\beta\eta} (M+\gamma G)_{-1}^{1-\eta}.$$

Equation (4) incorporates permanent displacement but not temporary displacement: the postponement of municipal construction either in the beginning or when major changes are made in the federal grants program. To measure temporary displacement, we enter as an explanatory variable the (real) budget authority for the program in each year less actual (real) outlays. This variable (A) should be large when the program really gets underway, but eventually should settle down to a steady state where, except for some years where major program changes occur, budget authority more or less equals outlays. Note that A is a flow rather than a stock since it represents a limited period of temporary confusion, and affects demand only during this period. Temporary displacement is added linearly so that an extra dollar will always have the same effect, regardless of the size of the unspent budget authority. This way of entering the variable is superior to the obvious alternative, to enter it multiplicatively. The multiplicative form would require that a doubling of this variable would always have the same percentage effect, whether the doubling is from a very small or very large base.

Adding temporary displacement and rearranging terms, we obtain the final form of the equation to be estimated:

$$(5) \quad M = \alpha^{\eta} H^{\beta\eta} (M+\gamma G)_{-1}^{1-\eta} - \gamma G - \theta A.$$

II. Data

To measure the extent and timing of the local response to federal expenditure, we use

[1] An alternative would be to adjust the price of sewer system construction to represent the federal share. This alternative, however, would only be appropriate if the community's entire demand for sewer system structures were matched with federal grants. In fact, only part of the demand is eligible for grants.

annual time-series data from 1949 to 1981. Time-series data are particularly useful here (as opposed to, say, a state cross section) because the program developed over time and changed sharply in some years.

The data series included the following: the stock of grants (where grants are measured as actual calendar year outlays of the Construction Grants Program), the stock of structures built with municipal funds (including a stock from expenditures necessary to match federal grants), and the residential housing stock. Each of the stocks of sewer system structures (federal and municipal) was composed of stocks of two components: sewer lines and treatment plants. All stocks were in 1972 dollars and were formed by the perpetual inventory method. Depreciation rates were 4 percent for lines, 10 percent for plants, and 2 percent for the residential housing stock. For plants and lines, depreciation rates are based on economic lifetimes provided by EPA. The spending series, which we cumulated into stocks for housing and sewer systems, were from U.S. Department of Commerce *Statistical Abstract...* (various issues) and *Value of New Construction...* (1975). Outlays and the budget authority of the Construction Grants Program were from U.S. Executive Office of the President, Office of Management and Budget (various issues) and the matching share and the spending on treatment plants and lines were from U.S. Executive Office of the President, Council on Environmental Quality (1976).

Since the construction of sewer systems began as far back as 1850, we began cumulating the stock then. Of course, most of this early stock had been depreciated by the beginning of the sample period. There was poor documentation of the split between plants and lines before 1965; we therefore treated all pre-1965 construction as lines.

III. Results of Estimation

Nonlinear least squares estimates of the demand for sewer system structures (equation (5)) are shown below. Estimates correcting for first-order serial correlation by means of a grid search were virtually identical.

Parameter	Estimate	t-value
$\alpha\eta$	.36	4.32
$\beta\eta$	.38	6.47
$1-\eta$	.60	9.98
γ	.67	11.92
θ	.28	6.52

Range: 1949–81; R^2: .998.

The coefficients of primary interest are those that measure permanent and temporary displacement (γ and θ). The estimate of .67 for γ, is strongly significant and is interpreted to mean that, for each dollar of federal expenditure, municipal expenditure is reduced by about two-thirds of $1 and, further, that each dollar of federal expenditure is worth 67¢ to the community. Of course, the total gain from the expenditure, including the gain to those outside the community, could be much larger; a primary rationale for legislation requiring wastewater treatment is that municipalities disregard externalities from untreated wastewater.

In addition to the permanent displacement, we also found evidence for temporary displacement. We estimated that, for each extra $1 of unspent budget authority, municipalities postpone about 28¢ of their expenditure. This temporary displacement is not trivial; the unspent budget authority was large at the beginning of the program.

The housing stock performs well as a scale variable.[2] Its coefficient is significant and the long-run coefficient turns out to be about .95 ($=\beta$), indicating that the dependence of the sewer system stock on the housing stock in the long run is almost exactly proportional.

IV. Further Results

To provide further information on the extent of displacement, we include several variations on equation (5) (see Table 1). The first variation (5a) replaces the variable A (tem-

[2] Because industrial activity can also generate water pollution, we tried as a scale variable a measure of the industrial capital stock added to the housing stock. We found, however, that the housing stock alone was superior.

TABLE 1—FURTHER RESULTS[a]

	Equation		
Parameter	(5a)	(5b)	(5c)
$\alpha\eta$	.48	1.10	.67
	(5.15)	(2.56)	(3.55)
$\beta\eta$	.32	.28	.36
	(6.21)	(4.02)	(6.62)
$1-\eta$	.65	.62	.56
	(12.23)	(7.48)	(6.31)
γ	.62	.49	1.14
	(11.24)	(5.20)	(8.46)
θ	1093.71	.22	.29
	(7.32)	(4.19)	(4.23)
Range: 1949–81			
R^2:	.999	.988	.996

[a] t-values are shown in parentheses.

porary displacement) with a simple dummy variable for 1972–74 to signify the beginning of the grants program.

The second variation (5b) changes the grant variable (G) from including only federal funds to including, as well, the local matching share. As a consequence, the discretionary stock now measures only the stock built with local funds other than those matching federal grants. This way of defining the M and G variables makes the entire cost of the project the source of displacement.

The third variation (5c) revises the measure of sewer system structures to include only sewer lines, not treatment plants. Of the two types of sewer system structures, treatment plants are more likely to generate benefits external to the community, and grants for treatment plants ought to involve less displacement than grants for lines.

In many ways, the variations confirm the qualitative results of the basic regression. Permanent displacement is always .5 or greater, and there is significant temporary displacement as well. Equation (5a), the dummy variable regression, shows almost the same permanent displacement as the basic regression. Equation (5b) has a smaller parameter for permanent displacement, but G is now correspondingly larger so that total (permanent) displacement is at least as large. Equation (5c), the equation for sewer lines only, shows displacement that is complete, which seems very reasonable since sewer lines are valued primarily for their local services; there is little of the externality issue here.

V. The Effect of the Program on Expenditures

Although the theory and empirical results are provided in terms of stocks, the effect of the grants program on expenditure flows (such as current expenditure on sewer system construction) is perhaps of greater interest. Consider as an example the situation in 1973. In 1973, grants were \$1.05 billion and unspent budget authority (the source of temporary displacement) was \$5.03 billion. Because permanent displacement was about 67 percent (from equation (5)), \$714 million was permanently displaced (out of \$1.05 billion in total grants), leaving a net addition to total spending of \$346 million. Since the temporary displacement parameter was about 28 percent of the \$5.03 billion in budget authority less grants, \$1.4 billion was temporarily displaced. In that year, the net add-on to sewer system construction from the program was therefore about −\$1.06 billion. Since the total spending that was displaced (including both permanent and temporary) was about \$2.1 billion, and there were about \$1.05 billion in grants, total displacement in 1973 was about 200 percent.

The year 1973 was unusual because the large federal expenditure had just started. In a later year, say 1980, grants were \$3.93 billion, and unspent budget authority was −\$525 million. The net permanent displacement was \$2.67 billion, permanent add-on was \$1.26 billion, and temporary displacement was −\$147 million. Net additions to spending were therefore \$1.407 billion and total displacement was about 64 percent (\$2.523 billion as a proportion of \$3.93 billion). In addition, we would expect some catch up from the temporary displacement of past years via the partial adjustment process.

VI. Summary and Conclusions

Evidence from time-series data suggests that EPA grants for construction of sewer systems displace municipal expenditure for the same type of project. The displacement involves temporary postponement of projects while grant applications are processed and permanent substitution of federal spending for what municipalities would have paid for themselves.

The finding of *temporary* displacement shows how grants may upset the spending plans of local government. Not only is there an administrative delay before expenditure can be made, but during the waiting period, municipalities may actually cut their own spending.

The finding of *permanent* displacement suggests that the grants do not increase the construction of sewer systems dollar for dollar; instead, state and local governmental units cut back their spending by about two-thirds.

That there is permanent displacement provides some evidence that the grants are valued by the recipients for the services they provide, not only for the income they bring to the region. To be more explicit, the findings do not accord with the usual assumption that the EPA is financing activities that the communities do not want. The way we are interpreting the parameters is standard. For example, Edward Gramlich and Harvey Galper (1978) interpret their parameters indicating displacement as a measure of how much federal grants contribute to the community's utility. In a different context, Timothy Smeeding (1977) uses displacement to measure the value to recipients of food stamps. As George Johnson and James Tomola (1977) point out, one likely objective of federal expenditures, even grants-in-aid, is pure revenue sharing, and this motive is best satisfied when there is 100 percent substitution of federal expenditures for state and local, that is, when grants have full value to recipients. Our evidence is that construction grants do have value to the recipients and, hence, have value as a program for partial revenue sharing.

Even so, the value of grants is eroded by the delay and by the administrative costs of preparing and evaluating grant proposals and of monitoring the grants. A question for future research is whether this erosion is offset by the efficiency gains of limiting externalities.

REFERENCES

Clarkson, Kenneth W., *Food Stamps and Nutrition*, Washington: American Enterprise Institute, April 1975.

Gramlich, Edward M. and Galper, Harvey, "State and Local Fiscal Behavior and Federal Grant Policy," *Brookings Papers on Economic Activity*, 1:1978, 191–216.

Johnson, George E. and Tomola, James D., "The Fiscal Substitution Effect of Alternative Approaches to Public Service Employment Policy," *Journal of Human Resources*, Winter 1977, *12*, 3–26.

Smeeding, Timothy M., "The Antipoverty Effectiveness of In-Kind Transfers," *Journal of Human Resources*, Summer 1977, *12*, 360–78.

West, David A. and Price, David W., "The Effects of Income, Assets, Food Programs and Household Size on Food Consumption," *American Journal of Agricultural Economics*, November 1976, *58*, 725–30.

U.S. Executive Office of the President, Council on Environmental Quality, Computer Printout, October 27, 1976.

______, **Office of Management and Budget,** *Special Analyses: Budget of the United States Government*, Washington: USGPO, various issues.

U.S. Department of Commerce, Bureau of the Census, *Statistical Abstract of the United States*, Washington, various issues.

______, *Value of New Construction Put in Place, 1947–1974* (Construction Reports-Series C30-74S), Washington: USGPO, 1975.

WHAT SURVIVES OF THE RATIONAL EXPECTATIONS REVOLUTION?

Rational Expectations and Macroeconomics in 1984

By ROBERT J. BARRO*

One of the cleverest features of the rational expectations revolution was the appropriation of the term "rational." Thereby, the opponents of this approach were forced into the defensive position of either being irrational or of modeling others as irrational, neither of which are comfortable positions for most economists. In fact, much of the rational expectations view—that expectations are formed sensibly given the information that people have (and are motivated to acquire)—has been generally accepted. This viewpoint has permanently and usefully altered the way that most macroeconomists build models and carry out evaluations of shifts in governmental behavior. In this sense the rational expectations revolution has triumphed decisively.[1]

The phrase, rational expectations macroeconomics, also suggests a particular theory of business fluctuations. This well-known theory shows how incomplete information about the quantity of money and the general price level can lead to nonneutrality of money. Specifically, changes in money lead to temporary confusions between general and relative prices, which lead in turn to adjustments of production and employment. Some extended versions of this model allow the real effects of monetary disturbances to persist over periods that are long enough to correspond to real world recessions and booms. There are also some intriguing implications for monetary policy—namely, the systematic part does not matter (aside from the inflation tax), and the erratic behavior tends to be harmful.

The rational expectations theory of business fluctuations and the empirical work that relates to this theory are surely interesting and suggestive. However, it seems a fair assessment that this research has not provided a definitive analysis of either monetary nonneutrality or of the business cycle more generally. Consequently, this work has received much criticism, some of which has even been insightful. But I should stress that the serious problems all arise in the attempt to explain the nonneutrality of money, which is surely the hardest problem in macroeconomics. Some recent research, which I discuss later, suggests that the solution to this problem may not be as crucial for the understanding of business cycles as many of us used to think.

One troublesome aspect is the place of rational expectations macroeconomics in the often political debate over Keynesian economics. At least implicitly, many people feel that what's bad for the rational expectations viewpoint is good for the Keynesian one, and vice versa. But it is hard to see how the problems in using the rational expectations approach to explain monetary nonneutrality can alleviate the theoretical and empirical shortcomings of the Keynesian model. This model is basically an incomplete theory that provides many prescriptions for activist macro policies, but which retains some serious inconsistencies with the rational behavior of individuals. Specifically, no one has been able to use elements such as information and mobility costs—which are obvious candidates for explaining the coordination problems of private markets—in order to

*Professor of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637.

[1] Even many Keynesian models now employ rational expectations, although amidst sticky prices and rationed quantities. Frankly, one has to wonder about the internal consistency of this procedure—as Laurence Weiss puts it, "If one is willing to posit 'reasonable' descriptions of how wages and prices are determined without reference to an explicit optimizing justification, then why not simply assert that expectations can be similarly modeled?" (1983, p. 6).

generate results that look Keynesian.[2] Similarly, the Keynesian model still has its difficulties with inflation and supply shocks—difficulties that were the prime reasons for the widespread and growing dissatisfaction with this model since the late 1960's.

If we look beyond the issue of monetary nonneutrality, then we do find areas of macroeconomics that use rational expectations and in which important recent progress has been made. One area concerns real theories of business fluctuations—that is, fluctuations in real economic activity that reflect underlying shocks to technology. Often people express skepticism that aggregate real shocks occur with sufficient size and frequency to account for a major part of the business cycle. For example, after mentioning the oil crises and harvest failures, one is often asked to name the third example of an important real shock. In this regard, I find David Lilien's research (1982) to be particularly promising. He shows that greater dispersion in the shifts to technology and tastes—with no necessary aggregate bias—can lead to significant and persistent effects on the aggregates of output and employment. In assessing the empirical significance of this approach, we could look at the major changes in patterns of international comparative advantage that have occurred since the early 1970's. These changes show up, for example, as a faster rate of decline in the share of U.S. output that is accounted for by the manufacturing sector. We would look also at the volatility in the relative prices of internationally traded goods, which are not limited to oil. These fluctuations in relative prices seem to have a great deal to do with the gyrations in real exchange rates over the last decade. Overall, it seems that real theories are especially promising for explaining the sharp fluctuations in real economic activity and the tendency for increases in unemployment rates since the early 1970's.

The other main objection to real business cycle theory is that it fails to address the link between money and real variables. At least through the 1930's, the main positive association between monetary aggregates and real variables derives from fluctuations in the quantity of financial intermediation—that is, the volume of deposits and loans—rather than from shifts in the monetary base. In many cases, the economic contractions were accompanied by banking panics, which tended to feature the suspension of convertibility between deposits and currency. As Ben Bernanke (1983) has stressed, it is not difficult to see why a sudden decline in the quantity of financial intermediation would have adverse real consequences for the economy. In fact, a cutback in financial intermediation is not so different from a negative shock to production functions, which is the type of disturbance that appears in real theories of business cycles. Thus, the main challenge is to explain why the earlier financial system was subject to occasional crises—not why these crises, once they occurred, would have serious repercussions on output and employment. It seems likely that deposit insurance plays an important role in this story.

Even in the post-World War II period, there is evidence that much of the interplay between money and real activity reflects fluctuations in deposits and in credit aggregates, rather than the monetary base. Thus, money may serve more as an indicator of changes in business conditions, rather than as a major exogenous influence on real variables. However, evidence on the interactions between the monetary base and real variables still suggests some amount of monetary nonneutrality, which we would like to explain. In fact, this type of interplay between monetary and real variables during relatively minor recessions may be quantitatively in line with the rational expectations models, which stress the role of incomplete information about money and prices. Thus, we may be able to resolve the puzzle of monetary nonneutrality by arguing first, that much of the empirical association between money and real variables is not evidence of nonneutrality, and second, that the existing theories can account for the relatively small amount of nonneutrality that remains.

[2] One of the more remarkable recent developments is the view that long-term contracts can rescue the old Keynesian case for policy activism. It is hard to see how the ability to contract could lessen the private economy's ability to deal with disturbances and thereby enhance the case for Keynesian macropolicies.

Another area in which important progress has been made concerns fiscal variables—that is, government expenditures, taxes and deficits. Here, the rational expectations viewpoint is consistent with real effects from some types of systematic macropolicies. For example, economic activity would generally respond to changes in the level or timing of government purchases and public services, as well as to the timing of taxes if these levies are not lump sum. However, the "Ricardian Theorem" says that choices between deficits and lump sum taxes do not matter for real variables. (They may or may not matter for the price level.)

There is empirical evidence that documents the expansionary effect on output from government purchases, especially for the sharp, temporary buildups that accompany wars. However, it is harder to verify the real effects from shifts in the timing of non-lump-sum taxes, probably because governments typically tailor their debt-management policies to avoid major swings in tax rates. In fact, this perspective leads to a useful positive theory of deficits—namely, they are increased by wars and recessions and, it turns out, by higher rates of expected inflation. This viewpoint explains U.S. deficits reasonably well since World War I, including the experience for 1982–83. Finally, there is little evidence that shifts in deficits are an independent source of business fluctuations (or of changes in interest rates). Thus, the Ricardian Theorem seems to be in good shape—and to have become remarkably respectable.

As to more progress, our understanding of macropolicy has been expanded particularly by Finn Kydland and Edward Prescott (1977), who brought out the essential distinction between rules and discretion. A rule provides commitments about future governmental behavior, for example, to honor patents on inventions or not to default on public debts. These commitments, which amount to a rule of law in the framing of governmental policy, can encourage efficient behavior of individuals, such as to invent things and to hold the government's debt. Thus, there can be a gain from the government's "tying its hands" in advance, rather than following the discretionary approach of optimizing at each date with the current state always taken as given (which might call for disallowing existing patents and defaulting on existing debts). This viewpoint has important insights for the desirability of rules in the context of monetary and fiscal policies. The gold standard and constitutional limits on money, taxes and spending can be viewed from this perspective as possibly useful rules. Further, if rules are absent, then the theory of discretionary policy provides predictions about the outcomes for inflation, monetary growth and other variables. In other words, we can derive a positive theory of governmental behavior that complements the usual theory of individual behavior. Proceeding this way, we can account for the high and variable inflation that prevails under the present "unruly" structure for monetary policy.

It is easy for me to appreciate the recent progress in macroeconomics since I have only to recall the level of knowledge that prevailed during my student days at Harvard. I remember vividly a lecture when I was a new graduate student in 1966. The opening speaker announced gleefully that the business cycle was dead. Then the main lecturer—on leave from Harvard at the Council of Economic Advisers—told us how the wonders of fine-tuning in macropolicy had been used virtually to eliminate recessions. (He even divided the gain in *GNP* by the number of economists in the United States to show how valuable each economist was, which made us students feel very warm and self-satisfied.)[3] No doubt we are sorry that the world does not work this way, but it must be progress to have found out.

The truth is that in the mid-1960's there was an artificial, essentially political consensus on the Keynesian model, which was not built on much supporting economic theory or empirical evidence. The evaporation of that consensus produced a letdown among many macroeconomists, not to mention policymakers and news reporters. But this breakdown was necessary in order for us to

[3] Probably there should be a distinction here between the average economist and the marginal economist. But that falls into the domain of price theory.

begin learning about the macroeconomy and to stimulate the development of superior methods of theoretical and empirical analysis. Over the last fifteen years we have generated a significant array of findings about the macroeconomy, and there is the promise of much more to come. Often people downgrade this progress because of a focus on the one area—namely the nonneutrality of money—in which the most problems have arisen. But overall we are in much better shape in macroeconomics—in terms of what we know and of knowing what we don't know—than we were fifteen years ago. Why, even the undergraduate textbooks in macroeconomics are getting better.

REFERENCES

Bernanke, Ben, Monetary Effects of the Financial Crises in the Propagation of the Great Depression,"*American Economic Review*, June 1983, *73*, 257–76.

Lilien, David, "Sectoral Shifts and Cyclical Unemployment," *Journal of Political Economy*, August 1982, *90*, 777–93.

Kydland, Finn and Prescott, Edward, "Rules rather than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, *85*, 473–91.

Weiss, Laurence, "Rational Expectations Models in Macroeconomics," unpublished, University of Chicago, 1983.

Efficiency and the Variability of Asset Prices

By STEPHEN F. LEROY*

It has now been a decade since the first of the variance-bounds papers was circulated in typescript. If initially less interest was displayed in this material than the authors had hoped and expected, the same is no longer true. This may be a good time to discuss a few of the many recent papers extending and criticizing the original results.

The central idea underlying the variance-bounds tests is very simple. Consider stock prices. The perfect foresight price of stock —that price which would prevail if future dividends x_{t+i} were known—is

$$(1) \qquad p_t^* = \sum_{i=1}^{\infty} \beta^i x_{t+i},$$

assuming a constant discount rate. Actual stock prices are assumed to be given by

$$p_t = E_t(p_t^*) \equiv E(p_t^* | I_t),$$

so that stock prices are the summed discounted value of expected dividends. The expectation is conditional on I_t, the information investors have at date t. Since p_t^* equals p_t plus a term which, under rational expectations, is orthogonal to p_t, we must have

$$(2) \qquad \mathrm{var}(p_t^*) \geq \mathrm{var}(p_t)$$

$$(3) \qquad \mathrm{var}(p_t^* | A) \geq \mathrm{var}(p_t | A)$$

for A any subset of I_t. To test the implication that p should be smoother than p^*, Sanford Grossman and Robert Shiller (1981), for example, calculated p^* from the backwards recursion $p_t^* = \beta(p_{t+1}^* + x_{t+1})$ (a consequence of (1)) and graphed it against p. By inspection, they argued, p^* is much less choppy than p, contrary to (2) and (3). Formal statistical tests performed by various authors supported the same conclusion.

Before reviewing possible criticisms of this reasoning, it is appropriate to discuss what conclusions follow from the apparent violation of the variance-bounds inequalities. This question of interpretation was handled very differently in the two original variance-bounds papers, a fact which appears to have escaped general notice. The dominant interpretation, that of Shiller, is that the results support the hypothesis that "stock prices are heavily influenced by fads or waves of optimistic or pessimistic 'market psychology'" (1981a, p. 294). To Shiller such an alternative "seems appealing, given the observed tendency of people to follow fads in other aspects of their lives and based on casual observation of the behavior of individual investors" (p. 298). He formulated a "fads" model and suggested that, if this alternative hypothesis were true, variance-bounds tests might be more likely to reject the incorrect efficient markets null hypothesis than would other types of market efficiency tests. This would explain why the variance-bounds tests reject the same model that conventional efficient-markets tests accept.

Others have followed Shiller in interpreting the variance-bounds tests as supporting the hypothesis of stock market irrationality. For example, Kenneth Arrow wrote that "A very rigorous analysis for the bond and stock markets [Shiller (1979), 1981b)] has shown the incompatibility of observed behavior with rational expectations models, at least in a simple form" (1983, p. 12), while Gardner Ackley in his presidential address to this Association a year ago observed that "Robert Shiller's recent paper [1981b] appears to demolish the possibility that movements of U.S. stock prices can be explained by the rational expectations of share holders" (p. 13).

It is remarkable how little evidence it takes to convince many economists, even some not otherwise sympathetic to magical explana-

*Department of Economics, University of California, Santa Barbara, CA 93106.

tions of economic phenomena, that the stock market is exempt from the same economic theory that explains the price of rutabagas. For many, stock market irrationality is an accepted fact, a matter of routine observation. Ackley, for example, went on to note that "Because stock prices are not fully rational, either in the large or even in the small, sharp-eyed members of several generations of my graduate students learned (not from me) to support themselves in reasonable comfort by playing on trivial systematic anomalies they had found in share price movements" (p. 14).

In contrast to the dominant interpretation, my article with Richard Porter (1981) regarded the violation of the variance-bounds inequalities as an anomaly, nothing more. Our conclusion was not that evidence had been provided in favor of some alternative to market efficiency, but merely that something was wrong either with a model that had been supported in other empirical work, or with statistical operations that seemed reasonable. If the various criticisms of the variance-bounds tests are accorded a more sympathetic interpretation here than in Shiller's work, that is because these criticisms support rather than conflict with LeRoy-Porter's conclusion, which was that something *must* be wrong with the tests, or the variance inequalities would not have been violated!

Criticisms of the variance-bounds conclusions fall into three groups:

1) *Econometric Problems.* The assumption of stationarity, or the correction made to achieve stationarity, has been criticized (Allan Kleidon, 1982, 1983; Terry Marsh and Robert Merton, 1983a,b). The fact that small-sample estimation problems strongly bias the variance-bounds tests toward rejection of market efficiency has been noted (Kleidon, 1982; Marjorie Flavin, 1983). Questions relating to statistical power have been raised (James Stock, 1982). It is not altogether clear that such statistical problems are sufficient to account for the dramatic rejections of the variance-bounds inequalities reported in the literature, but the authors just cited have provided convincing evidence that biases in the variance-bounds tests are much stronger than might have been expected.

2) *Wrong Model.* Like any test of market efficiency, the variance-bounds tests require specification of the model assumed to generate the data. In both the stock market and the bond market tests, an "expectations" model was adopted: my article with Porter employed the expected-present-value model in the former case, while Shiller assumed the expectations hypothesis in the latter. These models give accurate descriptions of asset prices only under restrictions which, if incorrect, will in some cases bias the variance-bounds tests toward rejection.

The expected-present-value representation of stock prices does not generally obtain in production economies under any specification of preferences, while it is generally true in exchange economies only under risk neutrality (myself, 1973, Robert Lucas, 1978; see my 1982a article for discussion). In an exchange economy, risk aversion generally increases stock price volatility relative to that of dividends (see my article with C. J. LaCivita, 1981; Ronald Michener, 1982; Grossman and Shiller), essentially because the more risk-averse individuals are, the more volatile asset prices must be to induce them to overcome their desire to smooth their consumption streams and instead consume the endowment. The same logic applies in at least some production economies. Therefore adoption of the present-value model biases the variance-bounds tests toward rejection of the efficient-market model to the extent that individuals are risk averse.

Matters are somewhat more complicated for volatility tests of the term structure of interest rates. With regard to real interest rates, the expectations hypothesis applies as an approximation if and only if real interest rates are nearly constant (see my 1982b, 1983 papers). In an exchange economy, this occurs if and only if individuals are nearly risk neutral (or there is nearly no nondiversifiable risk). However, since risk aversion increases the volatility of both long and short interest rates (rather than that of long relative to short), adoption of the expectations hypothesis does not bias the variance-bounds tests in any obvious way. In this respect the variance-bounds tests for interest rates differ from those for stock prices, which, as just shown, are biased toward rejection under risk aversion. In production economies, however, there is no general connection between risk aversion and interest rate volatility. For example, it is known that if individuals have constant relative risk aversion and production occurs under stochastic constant returns to scale subject to serially independent productivity shocks, then real interest rates of all maturities will be nonrandom. Hence if these specifications are approximately true, the expectations hypothesis will be approximately valid regardless of the extent of individuals' risk aversion.

With regard to nominal interest rates, few results are available (but see my 1984 article).

3) *Wrong Variance Measure*. It was observed that the evidence of violation of the variance-bounds inequalities which is most dramatic, particularly to those reluctant to venture far into econometric thickets, is the smoothness of p^* compared to p, apparently in contradiction to (2) and (3). In recent work (1982, 1983) Kleidon has questioned whether the smoothness of p^* really conflicts with market efficiency (the analysis to follow, although differing in detail from Kleidon's, draws from his papers the essential distinction between conditional and unconditional variances as measures of smoothness). In looking at Grossman-Shiller's graph, what is it that gives us the impression that p is choppier than p^*? Surely not the unconditional variance of p^* relative to p, since we have only a finite data interval. The variance restriction suggested by the smoothness of p^* relative to p is better represented by

$$(4)\qquad \operatorname{var}(p^*_{t+n} - p^*_t) < \operatorname{var}(p_{t+n} - p_t)$$

for n moderately small, or, alternatively,

$$(5)\qquad \operatorname{var}(p^*_{t+n} | p^*_t) < \operatorname{var}(p_{t+n} | p_t).$$

Now, if n were to approach infinity, these inequalities would both reduce to $\operatorname{var}(p^*_t) < \operatorname{var}(p_t)$, exactly the inequality which is contradicted by theory. Thus the question reduces to whether theory also contradicts (4) and (5) when n is small (note that (5) is not necessarily in conflict with (3) since p^*_t is not a function of I_t).

As it happens, for small n the inequalities (4) and (5) are exactly what should be expected in an efficient market if stocks are valued according to the expected present-value relation. To see this, suppose that dividends follow the first-order autoregressive process

$$(6)\qquad x_{t+1} = \rho x_t + \varepsilon_{t+1} \qquad |\rho| < 1,$$

where the ε_t may be taken to be normally distributed with zero mean and unit variance, and are serially independent. Then p^*_t and p_t can be expressed as functions of ρ, β, and all the ε, so we can evaluate the variances in (4) and (5) and verify that the observed inequalities are in fact generated for reasonable values of ρ and β, and small n.

To understand why p^* should be less choppy than p (in the sense that the variance of $p^*_{t+n} - p^*_t$ is less than that of $p_{t+n} - p_t$ for small n), note that when p^*_t is expressed as a function of the past, current and future shocks to dividends, the weights are very smooth, particularly when β and ρ are near one (the greatest coefficient is that of ε_{t+1}; the coefficients of future ε decline geometrically with factor β, while those of past ε decline geometrically with factor ρ). Hence for small n the weights on each of the ε in the expression for $p^*_{t+n} - p^*_t$ is very small, implying that $\operatorname{var}(p^*_{t+n} - p^*_t)$ will be small. With p_t, however, the weights on all future ε are zero. Consequently, $\varepsilon_{t+1}, \dots, \varepsilon_{t+n}$ contribute to p_{t+n}, but not to p_t, and therefore their coefficients in the expressions for $p_{t+n} - p_t$ are much greater than in that for $p^*_{t+n} - p^*_t$.

How important quantitatively is this effect? To answer this question, the variances in question were calculated for one value of β (.9; the results are not sensitive to the value chosen for β), two values of ρ (.8 and .99) and a range of values for n. By inspection of Table 1, the story is the same whether we compare $\operatorname{var}(p^*_{t+n} - p^*_t)$ with $\operatorname{var}(p_{t+n} - p_t)$ or $\operatorname{var}(p^*_{t+n} | p^*_t)$ with $\operatorname{var}(p_{t+n} | p_t)$, so we consider the former comparison only. For $\rho = .8$, the unconditional ($n = \infty$) variance of $p^*_{t+n} - p^*_t$ is four times as great as that of $p_{t+n} - p_t$, but for differencing intervals of one or two periods, the conclusion would be that p^* is smoother than p.

TABLE 1—"SMOOTHNESS" OF p^* RELATIVE TO p

	(1)	(2)	(3)	(4)
$\rho = .8$	$\text{var}(p_{t+n} - p_t)$		$\text{var}(p_{t+n} \mid p_t)$	
n		$\text{var}(p^*_{t+n} - p^*_t)$		$\text{var}(p^*_{t+n} \mid p^*_{t+n})$
1	7	2	7	2
2	13	6	11	6
5	25	26	16	24
20	36	114	18	69
50	37	144	18	73
100	37	145	18	73
∞	37	145	18	73
$\rho = .99$				
n				
1	67	4	67	4
2	134	15	132	15
5	329	82	321	82
20	1223	809	1112	765
50	2653	2467	2129	2058
100	4257	4426	2908	3108
∞	6716	7433	3358	3717

Note:
Col. 1: $\text{var}(p_{t+n} - p_t) = [(1-\rho^{2n}) + (1-\rho^n)^2](\beta\rho)^2 / [(1-\beta\rho)^2(1-\rho^2)]$.
Col. 2: $\text{var}(p^*_{t+n} - p^*_t) = \beta^2[\Sigma^n_{i=0}(\beta^i - \rho^{n-i})^2 + (1-\rho^n)^2\rho^2/(1-\rho^2) + (\beta^n - 1)^2\beta^2/(1-\beta^2)]/(1-\beta\rho)^2$.
Col. 3: $\text{var}(p_{t+n} \mid p_t) = (1-\rho^{2n})(\beta\rho)^2 / [(1-\rho^2)(1-\beta\rho)^2]$.
Col. 4: $\text{var}(p^*_{t+n} \mid p^*_t) = \gamma_0 - (\gamma_n)^2/\gamma_0$,
where $\gamma_n = \text{cov}(p^*_t, p^*_{t+n}) = \beta^2[\Sigma^n_{i=0}\beta^i\rho^{n-i} + \beta^{n+2} / (1-\beta^2) + \rho^{n+2}/(1-\rho^2)]/(1-\beta\rho)^2$.

The greater smoothness of p^* relative to p for small n is much more dramatic for $\rho = .99$, so that dividends are almost a random walk. Many would regard .99 as more plausible on empirical grounds than .8. As before, we verify that $\text{var}(p^*_{t+\infty} - p^*_t) > \text{var}(p_{t+\infty} - p_t)$, as the variance-bounds theory requires. But for n as high as 50, p^* is actually smoother than p, and the difference is of orders of magnitude for $n = 1$ or 2! Those who, like Kleidon, find these assumed parameter values realistic are therefore hardly surprised that Grossman-Shiller's p^* looks smoother than their p, even for a century of data. Indeed, Kleidon sees in the greater smoothness of p^* relative to p a confirmation of market efficiency (given the maintained assumption of a constant discount rate) rather than a contradiction.

In sum, the recent discussions have offered a number of independent possible explanations for the violations of the variance-bounds inequalities. As yet, no smoking gun —no single culprit unambiguously responsible for violation of the variance-bounds inequalities—has been found. However, our position may well be that of Hercule Poirot in Agatha Christie's *Murder on the Orient Express*, who found that all the suspects were guilty of the murder. It is appropriate to conclude, with Marsh and Merton, that the burden of proof is now on those who contend that asset prices are too volatile, rather than on those who view the observed behavior of asset prices as consistent with market efficiency.

REFERENCES

Ackley, Gardner, "Commodities and Capital: Prices and Quantities," *American Economic Review*, March 1983, *73*, 1–16.

Arrow, Kenneth J., "Behavior Under Uncertainty and Its Implications for Policy," Technical Report No. 399, IMSSS, Stanford University, 1983.

Flavin, Marjorie, "Excess Volatility in the Financial Markets: A Reassessment of the Empirical Evidence," *Journal of Political Economy*, December, 1983, *91*, 929–56.

Grossman, Sanford J. and Robert J. Shiller, "The Determinants of the Variability of Stock Market Prices," *American Economic Review Proceedings*, May 1981, *71*, 222–27.

Kleidon, Allan W., "Bias in Small Sample Tests of Stock Price Rationality," reproduced, Stanford University, 1982.

______, "Variance Bounds Tests and Stock Price Valuation Models," reproduced, Stanford University, 1983.

LeRoy, Stephen F., "Risk-Aversion and the Martingale Property of Stock Prices," *International Economic Review*, June 1973, *14*, 436–46.

______, (1982a) "Expectations Models of Asset Prices: A Survey of Theory," *Journal of Finance*, March 1982, *37*, 185–217.

______, (1982b) "Risk-Aversion and the Term Structure of Real Interest Rates," *Economics Letters*, No. 3-4, 1982, *10*,

355–61.

______, "Risk-Aversion and the Term Structure of Real Interest Rates: Correction," *Economics Letters*, No. 3-4, 1983, *12*, 339–40.

______, "Nominal Prices and Interest Rates in General Equilibrium: Endowment Shocks," *Journal of Business*, April 1984.

______ **and LaCivita, C. J.**, "Risk-Aversion and the Dispersion of Asset Prices," *Journal of Business*, October 1981, *54*, 535–47.

______**and Porter, Richard D.**, "The Present-Value Relation: Tests Based on Implied Variance Bounds," *Econometrica*, May 1981, *49*, 555–74.

Lucas, Robert E., Jr., "Asset Prices in an Exchange Economy," *Econometrica*, November 1978, *46*, 1429–45.

Marsh, Terry A. and Merton, Robert C., (1983a) "Earnings Variability and Variance Bounds Tests for Stock Market Prices: A Comment," reproduced, Massachussetts Institute of Technology, 1983.

______ **and** ______, (1983b) "Aggregate Dividends Behavior and Its Implications for Tests of Stock Market Rationality," reproduced, Massachussetts Institute of Technology, 1983.

Michener, Ronald W., "Variance Bounds in a Simple Model of Asset Pricing," *Journal of Political Economy*, February 1982, *90*, 166–75.

Shiller, Robert J., "The Volatility of Long-Term Interest Rates and Expectations Models of the Term Structure," *Journal of Political Economy*, December 1979, *87*, 1190–219.

______, (1981a) "The Use of Volatility Measures in Assessing Market Efficiency," *Journal of Finance*, May 1981, *36*, 291–304.

______, (1981b) "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?," *American Economic Review*, June 1981, *71*, 421–36.

Stock, James H., "Tests of Market Efficiency When Consumers are Risk Averse," reproduced, University of California-Berkeley, 1982.

Interpreting the Statistical Failures of Some Rational Expectations Macroeconomic Models

By JULIO J. ROTEMBERG*

This paper attempts to gauge the economic importance of the statistical rejections of empirical rational expectations models by concentrating on two examples; the first is concerned with aggregate consumption, the second with labor demand.

One of the principal advantages of rational expectations macroeconomic models over previous ones is that the former are subject to meaningful statistical tests. These tests typically check whether aggregate data contain correlations which the model doesn't predict. In prerational expectations macroeconomic models, it is relatively straightforward to amend the model to take into account the correlations present in the data. This is much harder to do in the context of rational expectations models. In many of these models the estimated parameters are parameters of the objective functions of economic agents. These parameters are estimated by fitting these agents' reaction functions. When correlations are present that are not predicted by the specific objective function under consideration, it is difficult to know how to parsimoniously change the objective function. So, it isn't too surprising that the first generation of published empirical rational expectations models have been statistically rejected by the data to which they have been applied. These models are extremely simple in many dimensions, so it is undoubtedly too early to predict the demise of the rational expectations research program on the basis of these failures. On the other hand, it is important to discuss whether all we can learn from these empirical efforts is that different (and presumably more complex) models are needed.

Rational expectations models attempt to estimate parameters which are not subject to the critique of Robert Lucas (1976). Such parameters would remain invariant to changes in economic policy. It is often asserted that the parameters describing the objective functions of economic agents fulfill this requirement. This may be problematic insofar as the election of those who change economic policy depends precisely on the objectives of economic agents. It is probably more appropriate to view parameters to be free of the Lucas criticism when they can be used for certain conceptual experiments.

One such conceptual experiment asks what the effect of a typical monetary shock is in the context of an unchanged economic policy regime. For this experiment the entire set of covariances between money and other economic variables may well be free of the Lucas criticism. One can probably safely assume that these covariances aren't affected by typical shocks.

Another conceptual experiment asks how consumption would respond if preferences were unchanged, but the real interest rate became permanently higher. To answer this question, one would ideally want to know the current utility functions of consumers. Similarly, to discover how much employment would rise if output or the real wage were permanently higher, one would want to know the firm's cost or profit functions. So researchers have tried and continue to try to obtain the parameters of cost, profit, and utility functions. These efforts have generally not managed to account for all the correlations present in the aggregate data under study. However, insofar as these data all suggest similar long-run responses of consumption to interest rates or of employment to output, they shed light on conceptual experiments of interest. Simple, statistically rejected models can be viewed as reasonable

*Massachusetts Institute of Technology, E52-434, Cambridge, MA 02139. I thank Olivier Blanchard, Lars Hansen, Jim Poterba, and Lawrence Summers for helpful discussions.

approximations if the data suggest that the parameters of interest lie on a relatively small subset of the parameter space.

In this paper, I focus only on models of tastes and technology which are estimated by instrumental variables. This estimation technique is ideally suited for the estimation of rational expectations models. These models almost invariably include equations in which current behavior depends on the mathematical expectation of future variables. The residuals created by replacing these expectations with the actual values of future variables thus can be interpreted as forecast errors. Forecast errors must, by definition, be uncorrelated with variables whose value is currently known. Such variables make ideal instruments.

Suppose there are m instruments, one equation, and k parameters. Lars Hansen (1982) shows that, if the model is correctly specified, consistent estimates are obtained by taking k linear combinations of the m inner products of instruments and residuals, and setting them to zero. These k linear combinations are given by a $k \times m$ weighting matrix. Hansen also provides a test of the model which can be applied when m exceeds k. The failure of this test has various interpretations. First, it means that the instruments are, in fact, correlated with the residuals. Thus these are not just the forecast errors of rational agents. Either the agents don't forecast rationally, or, more appealingly, the model is incorrectly specified. This misspecification, in turn, can be of two types. First, the model might simply be wrong. The true utility function or production function might have a different functional form, or it might depend on additional variables. Second, the equation's error term may be due in part to randomness of preferences or technology. This component of the error term might be serially correlated as well as being correlated with the instruments. Another interpretation, stressed by Jerry Hausman (1978) for a test, is shown by Whitney Newey (1982) to be often equivalent to Hansen's in that different instruments lead to different estimates. In my 1983 paper, I show that when the model fails these specification tests, one can obtain arbitrary estimates by varying the weighting matrix. This is particularly damaging because, when the model is misspecified, traditional methods for selecting a weighting matrix cease to be applicable. There is no reason to prefer two-stage least squares over any other weighting matrix. This could be interpreted as saying that if the model is misspecified even slightly so that estimates obtained using different combinations of k instruments are arbitrarily close to each other, the data nonetheless are consistently with *any* parameter vector. However, my paper also shows that if one is willing to assume that the misspecification has certain untestable properties, some of this bleakness is lifted. In particular, suppose one isn't willing to assert that the means of the products of instruments and residuals are equal to zero with probability one. Instead, one is willing only to assume that one's subjective distributions over these means have mean zero. These subjective distributions also have nonzero variances since one isn't sure the model is correctly specified. Then suppose the subjective covariance of the mean of the product of one instrument and the residuals with the mean of the product of another instrument and the residuals is zero. In other words, knowledge of the misspecification caused by one instrument doesn't convey information on the misspecification caused by another. This may be a reasonable set of beliefs particularly when the misspecification is due to random variations in preferences or technology. Moreover, in this case the parameters themselves still have economic meaning even though the model is misspecified. One might thus want to obtain the vector of parameter estimates which is, on average, least polluted by the correlation between residuals and instruments. My earlier paper shows that the parameter vector which in some sense minimizes the effect of misspecification is strictly inside the range of the estimates obtained by using all combinations of only k instruments. This parameter vector minimizes the asymptotic covariance matrix of the misspecified estimates around the true parameters. This suggests analysis of this range when the model fails a specification test. This is intuitively appealing since the failure of the model is due to the fact that different instruments lead to different estimates. One would like to know how different, from an economic point of view, these estimates actually are. If this range is large, then the data are unable to locate even the parameter vector which keeps the effect of misspecification small. Otherwise the region of parameters which could be used for conceptual experiments is small.

It is important to note at the outset that the study of this range presents difficulties in small samples. This is so because, in such samples, different combinations of instruments give rise to different estimates even when the model is correctly specified. In this case, of course, these differences vanish asymptotically. The failure of specification tests asserts that, in some sense, the range of estimates is bigger than it would have been if the model were correct. Moreover, this range doesn't vanish asymptotically for these models. My earlier results concern this asymptotic range, which unfortunately is not just imprecisely estimated, but also probably overstated in small samples.

In this paper I study some of these small sample ranges. I focus on those which obtain for a simple model of consumer preferences. This model is based on the work of Hansen and Kenneth Singleton (1982). Then I analyze some intertemporal cost functions which incorporate the cost of changing employment. These are inspired by John Kennan (1979). The ranges in both cases are fairly large.

I. Consumption

Per capita consumption is assumed to be given by the consumption of a single infinitely lived individual who maximizes the expected value of an additively separable utility function. This utility function has constant relative risk aversion so the individual maximizes at t:

$$(1) \quad E_t \sum_{\tau=t}^{\infty} \rho^{\tau-t}(C_\tau^{1-\gamma}-1)/(1-\gamma)$$

where E_t takes expectations conditional on information at time t, C_τ is consumption at τ, ρ is the discount factor while γ is the index of relative risk aversion. The individual also holds some assets in positive quantities. By giving up one unit of consumption at t, the individual gets P_t dollars at t. Investing these in the asset, he receives $P_t(1+r_t)$ dollars at $t+1$. These are sufficient to purchase $P_t(1+r_t)/P_{t+1}$ units of consumption at $t+1$. If the individual maximizes utility, he can't be made better off by holding either more or less of the asset. So

$$(2) \quad C_t^{-\gamma} = E_t\rho(P_t(1+r_t)/P_{t+1})C_{t+1}^{-\gamma},$$

where the left-hand side of (2) is the loss in utility from giving up one unit of consumption at t. Hence:

$$(3) \quad \rho(C_{t+1}/C_t)^{-\gamma}P_t(1+r_t)/P_{t+1} = 1+\varepsilon_{t+1}$$

where ε_{t+1} has mean zero conditional on information available at t. Taking logarithms on both sides and approximating $\ln(1+\varepsilon_{t+1})$ by the first two terms of the Taylor expansion,

$$(4) \quad \phi + \ln P_t(1+r_t) \\ /P_{t+1} - \gamma\ln(C_{t+1}/C_t) = \varepsilon_{t+1},$$

where ϕ is the logarithm of ρ plus half of the variance of ε_{t+1} which is assumed to be constant. Estimates of γ and ϕ can be obtained by instrumental variables applied in equation (4).

I use seasonally adjusted per capita nondurable consumption for C_t, the nondurable consumption deflator for P_t, and experiment with various different returns for r_t. In particular I consider the after-tax return on the S&P 500. This return is constructed by applying different tax rates on dividends and on capital gains as explained in my paper with James Poterba (1983). I also consider the after-tax return on U.S. Treasury bills and on savings accounts. The data are quarterly and extend from the third quarter of 1955 to the first quarter of 1982. The models are first fitted with two-stage least squares using the following seven instruments: a constant, $\ln(C_t/C_{t-1})$, $\ln(C_{t-1}/C_{t-2})$, $\ln(Y_t/Y_{t-1})$, $\ln(Y_{t-1}/Y_{t-2})$, $\ln(1+r_{t-1})P_{t-1}/P_t$, and $\ln(1+r_{t-2})P_{t-2}/$

P_{t-1}. Here Y_t is per capita disposable income at t. The model is rejected at the 1 percent level when after-tax Treasury bills and after-tax savings accounts are used but only at the 10 percent level with stock returns. Such rejections throw doubt on the usefulness of the parameter estimates.

The main parameter of interest is γ. The inverse of γ is the intertemporal elasticity of substitution (i.e., the elasticity of C_{t+1}/C_t with respect to the return $(1+r_t)P_t/P_{t+1}$). Also a low γ implies relatively little risk aversion, a strong substitution effect, and thus a tendency for savings to rise as the real return rises permanently. To see whether the data are relatively unanimous on γ, I fit the 21 exactly identified models which use only two of my seven instruments. When I use the after-tax rate of return on equity, γ ranges from -6.42 to 6.85. Negative γs imply a convex utility function. Such a utility function is on a priori grounds inconsistent with the data. It implies that the utility-maximizing agent would consume only zero in many periods. So one might be interested only in the positive γs. The smallest of these is .56 when the equity return is used. When I use the return on Treasury bills, positive γ's range from .14 to 12.76. Finally, when I use the after-tax return on savings accounts, they vary from .13 to 8.45. While these are undoubtedly very large regions, they have one thing in common. Namely, they do not include very high γs. These aggregate data point to only moderate risk aversion. This fact may well be useful in dynamic simulations of simple general equilibrium models.

On the other hand, the model itself needs to be changed substantially to even account for the correlations present in the aggregate data. First, more goods must be included in (1). However, simple attempts to include leisure in (1) lead to failure as reported by Gregory Mankiw, myself, and Lawrence Summers (1983). This failure comes from the inability to obtain reasonable parameter estimates. Second, the model must be modified to explain why rate of return dominated assets like money are willingly held. Some progress in this direction is reported in Poterba-Rotemberg. Third, the correlation between consumption and income suggests that some of consumption is carried out by liquidity constrained individuals. Thus modifying the model to take into account some individual heterogeneity appears promising.

II. Labor Demand

A standard view is that it is costly for firms to instantaneously adjust the factors that they hire. In the absence of these costs of adjustment they would hire N_t^* employees at t. If they have a different number of employees, their costs are higher. For larger employment, these additional costs are in the form of larger wage bills. For lower employment, the firm must ask its workers to work overtime, thus raising costs and reducing efficiency. On the other hand, changing employment is costly due to the need to pay training and severance pay. Thus a firm may minimize the expected value of total employment related costs which are given by

$$(5)\quad E_t \sum_{\tau=t}^{\infty} R_{t,\tau}(n_\tau - n_\tau^*)^2 + \beta(n_\tau - n_{\tau-1})^2,$$

where n_t is the logarithm of N_t while $R_{t,\tau}$ is the discount factor applied at t to costs incurred at τ. For a firm to be minimizing costs, the derivative with respect to n_t in (5) must be zero:

$$(6)\quad n_t - n_t^* + \beta(n_t - n_{t-1}) - E_t R_{t,t+1}\beta(n_{t+1} - n_t) = 0.$$

By increasing n_t, the firms induces $(n_t - n_t^*)$ costs of being away from n_t^*, incurs $\beta(n_t - n_{t-1})$ adjustment costs, but reduces the expected value of future adjustment costs by the $E_t R_{t,t+1}\beta(n_{t+1} - n_t)$. The sum of these additions to costs must be zero. Equation (6) can be estimated once n_t^* is specified. Kennan assumes that n_t^* is proportional to the logarithm of total production q_t. As in Kennan, I apply the model separately to the total manufacturing of durables and nondurables. The industrial production of these sectors provides the measure of q_t. The data on employment and production are seasonally adjusted, and both a linear trend and the mean of the variables have been removed. I use the after-tax nominal return on equity as the nominal rate of interest by which future costs are discounted. To obtain the real discount factor $R_{t,t+1}$, I use the wholesale price index of the corresponding sector. So $R_{t,t+1}$ in nondurables is given by $P_{t+1}^N/[(1+r_t)P_t^N]$ where P_t^N is the price index for nondurables at t and r_t is the rate of return on equity. The discount factor for durables is constructed analogously. These considerations allow (6) to be rewritten as a regression equation.

$$(7)\quad n_t-\alpha q_t+\beta\left[(n_t-n_{t-1})-R_{t,t+1}(n_{t+1}-n_t)\right]=\varepsilon_{t+1},$$

where ε_{t+1} has mean zero conditional on information available at t while α is the long-run elasticity of employment with respect to production.

I first estimate this equation by two-stage least squares using data from the third quarter of 1954 to the first quarter of 1982. The equation is rejected in both sectors with 99 percent confidence when I use as instruments the detrended values at t and $t-1$ of employment, production, and real hourly earnings in each sector.

I then estimate for each sector the 15 models which use only two of these six instruments. The resulting polyhedron of estimates is quite large. For nondurables, α ranges from .03 to 6.84 while β ranges from -56.71 to 383.35. Similarly for durables, α varies between -1.44 and 5.41 while β varies between -48.16 and 249.98. Once again, negative values of α and β can be dismissed on a priori grounds. Negative αs mean that more detrended output is ideally produced with fewer employees. Negative βs make it optimal to induce large fluctuations in employment from period to period. Such negative values would thus have zero likelihood if the model were completely specified. Even so, the smallest positive βs are 3.88 for durables and 2.21 for nondurables. The smallest positive α for a set of estimates with positive β is .69 for durables and .72 for nondurables. The long-run elasticity of employment with respect to output varies by a factor of ten starting at about 2/3. Instead β, the ratio of marginal costs of adjustment to costs of being away from n^*, varies by a factor of 100. To see the significance of this variation it suffices to calculate how long it would take n to reach 1/2 of the adjustment to its long-run value after a once and for all increase in q_t. To make this calculation, I assume $R_{t,t+1}$ is constant and equal to .99. Then it takes less than two quarters to complete half the adjustment if β is 5. On the other hand, if β is 300, half the adjustment takes over three years. So there is substantial uncertainty surrounding the economic importance of these costs of adjustment.

A more general model would include other inputs such as capital and its costs of adjustment. Then, n^* depends on output, on the costs of various inputs, and on the level of other quasi-fixed factors. Robert Pindyck and I (1983) present such a model and find results which are much more consistent across instrument lists. However, we do not analyze as many lists of instruments as are considered here. Incidentally, we find rather small costs of adjusting labor.

Finally, insofar as the misspecification is due to the presence of taste shocks in (4) or technology shocks in (7), these need to be modeled explicitly to discuss which instruments remain valid.

III. Conclusions

I have shown that for some very simple macroeconomic models, the statistical rejections to which they are subject have economic meaning. The rejections mean that different parts of the data lead to different estimates. The economic meaning of these rejections is that the difference between the various estimates is large. This makes it hard to pin down even the parameter vector which is least affected by misspecification. However, in the case of the consumption model, the data consistently suggest that the parameter of relative risk aversion is fairly small.

It is hoped that along the road towards models which actually account for all the correlations present in aggregate data, some models will be found which are better in the following sense. While these models will fail specification tests, the region of parameters which different data point to will be small enough to allow us to make some meaningful economic inferences.

REFERENCES

Hansen, Lars R., "Large Sample Properties of Method of Moments Estimators," *Econometrica*, July 1982, *50*, 1029–54.

______ **and Singleton, Kenneth J.,** "Generalized Instrumental Variable Estimation of Nonlinear Rational Expectations Models," *Econometrica*, September 1982, *50*, 1269–87.

Hausman, Jerry A., "Specification Tests in Econometrics," *Econometrica*, November 1978, *46*, 1251–72.

Kennan, John, "The Estimation of Partial Adjustment Models with Rational Expectations," *Econometrica*, November 1979, *47*, 1441–56.

Lucas, R. E., Jr., "Econometric Policy Evaluation: A Critique," in Karl Brunner and Allan Meltzer, eds., *The Phillips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl., 1976, 19–46.

Mankiw, G. N., Rotemberg, J. J., and Summers, L. H., "Intertemporal Substitution in Macroeconomics," mimeo., 1983.

Newey, Whitney K. "Generalized Method of Moments Specification Testing," mimeo., 1982.

Pindyck, Robert S. and Rotemberg, Julio J., "Dynamic Factor Demands and the Effect of Energy Price Shocks," *American Economic Review*, December 1983, *73*, 1066–79.

Poterba, James and Rotemberg, Julio J., "Money in the Utility Function: An Empirical Implementation," mimeo., 1983.

Rotemberg, J. J., "Instrumental Variable Estimation of Misspecified Models," mimeo., 1983.

Informational Imperfections in the Capital Market and Macroeconomic Fluctuations

By BRUCE GREENWALD, JOSEPH E. STIGLITZ, AND ANDREW WEISS*

Traditional neoclassical theory has one clear, unambiguous, and verifiable prediction: all factors which have a positive price are fully utilized. In recent years, there have been several responses to the apparent inconsistency between the predictions of neoclassical theory and what has in fact been observed. The first is to deny the empirical observations: the 25 percent of the population that were unemployed in the Great Depression, let alone the 10 percent of the population that were unemployed in the Reagan recession, were not involuntarily unemployed. This seems to us, at best, semantic quibbling, and we shall have nothing further to say here concerning that view. The second is to argue, without much justification, that there are two regimes; traditional neoclassical theory applies in "normal times." It seems more plausible to us that the market failures represented by the Great Depression are always present in the economy, but difficult to detect; it is only when they reach the proportions that they do periodically that we can no longer ignore them.

A third approach is to modify the standard theory, to assume that wages and prices are fixed. This approach has rightfully been criticized both for its *ad hocery* and its inconsistency—why should rational profit-maximizing firms, obeying all of the other neoclassical assumptions, not cut their prices in the face of excess demand?

This paper is part of an attempt to develop a consistent set of micro foundations for macroeconomics, based on imperfect information. We focus here on the capital market. Keynes argued that the sharp drop in investment and the failure of the interest rate to fall sufficiently to restore investment to a normal level was a central part of the description of any business cycle. Keynes' analysis of investment was, however, basically a neoclassical analysis: it was the failure of the real interest rate (the long-term bond rate) to fall sufficiently that was the source of the problem.

Three aspects of this analysis have always been troubling: first, Keynes' explanation of the failure of real interest rates to fall, the liquidity trap, is not persuasive. Second, surveys suggest that firms' investment behavior is not particularly sensitive to the interest rate that they pay. Third, it has always seemed difficult to account for the magnitude of the fluctuations in investment in terms of the observed magnitudes of variations in real interest rates, outputs, wages, and prices, unless firms are very risk averse; and it is hard to reconcile high degrees of risk aversion on the part of firms with well-functioning (neoclassical) capital markets.

This paper is based on the hypothesis that Keynes' judgment concerning the importance of fluctuations in investment is correct, but that he incorrectly analyzed the determinants of investment behavior. We argue that:

1) Many firms face credit constraints; thus it is the availability of credit, not the price which they have to pay, which restricts their investment, or when it is working capital which is curtailed, which limits their production.

2) Firms that are not credit constrained may still face an increase in the *effective* cost

*Greenwald: Harvard Business School, Boston, MA 02163; Stiglitz: Princeton University, Princeton, NJ 08544; Weiss: Columbia University, New York, NY 10027. Financial support of the National Science Foundation is gratefully acknowledged.

of capital, which induces them to reduce their investment. (The increase in the effective cost of capital has further effects, for example, on the pricing decisions of firms.)

I. The Debt Market

The main informational problem facing banks is that they do not know how the money they lend is being invested. Stiglitz-Weiss (1981, 1983) showed that an increase in the interest rate charged borrowers will, in general, increase the average riskiness of the projects a bank is financing. This is either because borrowers switch to riskier projects, or because safer projects become relatively less attractive and so investors with safe projects do not apply for loans. The effect on the riskiness of loans may outweigh the direct gain to the bank from increasing its interest rate. Thus, the bank's profit may be maximized at an interest rate at which there is an excess demand for loanable funds.

This kind of phenomenon (an interior price maximum and rationing, which may also occur in the labor market) helps to explain business cycles in three ways. First, and most obviously, it provides a rationale for the persistence of non-market-clearing. Second, it may account for variations in a firm's cost of capital which are unrelated to observed variations in interest rates. The likelihood and severity of credit rationing may well increase in a recession without necessarily any concurrent change in interest rates. An increase in credit rationing might be expected both because of greater uncertainty concerning the prospects of firms, and an increase in the deadweight loss associated with bankruptcy. Third, information-based rationing models can explain how stabilization policy is likely to work. For example, monetary policies which seek to increase investment by lowering interest rates will not have the desired effect: there is no shortage of willing borrowers. However, policies that increase the availability of loanable funds will increase investment, even though they may not affect the level of interest rates at all.

There are two objections to our credit rationing theory as an explanation of the cyclical fluctuations in investment. First, why don't firms that face credit constraints from banks attempt to raise capital by some other means, in particular, by issuing new equity. And second, many firms that do not appear to be credit constrained also seem to reduce their investment dramatically.

Thus, a necessary complement to the theory of credit rationing is a theory of informational imperfections in equity markets. This we present in the next two sections.

II. Equity-Markets

A firm's ability to raise equity capital is limited by informational imperfections for two basic reasons. First, incentive problems may intensify when a firm is equity financed. Managers, who receive only a small fraction of any additional profit, are likely to put forth less-than-optimal amounts of effort. Imposing large bankruptcy costs on managers may act as a spur to added effort and the value of these incentives is reduced by additional equity finance. Debt financing also allows managers less flexibility in disposing of net income than equity does. Thus, equity funds may reduce the value of a firm by allowing more "profit" to be diverted to the private uses of the firm's managers. Finally, lenders have the power to discipline managers by withdrawing their funds. This is a sanction which can be imposed piecemeal, and may therefore be more effective than share voting to which majority rule applies.[1]

Second, signalling effects may restrict a firm's access to equity markets. Managers of firms, which they know to be "good," may be willing to assume greater debt burdens. Both the absolute level of bankruptcy risk and any incremental increase due to added debt will be smaller for good than for "bad" firms. Greater reliance on debt by good firms means that equity will predominantly be sold by inferior ones (see Stephen Ross, 1977). Thus, attempting to sell equity may convey a strong negative signal about a firm's quality and reduce its market value accordingly.

[1] There are well-known impediments, both theoretical and practical, to shareholder control whether mediated via takeovers or normal corporate governance.

The model presented in this paper analyzes the cyclical cost of capital implications of the signalling process just described as an example of the macroeconomic impact of the many limitations on equity issue which are noted above. It provides an explanation for large, but not directly observable, variations in the marginal cost of capital (to be distinguished from the average cost of capital measured for example by Tobin's q) which can account for many of the variations in investment which are commonly associated with business cycles. The negative signal associated with issuing equity means that the cost of equity is prohibitive for many firms. Thus, the effective marginal cost of capital is the marginal cost of debt which consists of the monetary cost of interest plus the marginal increase in expected bankruptcy cost associated with additional debt. The latter bankruptcy cost will increase as a firm faces unexpectedly adverse economic conditions, and may do so dramatically. Moreover, it is likely that the adverse signal associated with issuing equity will intensify and place equity finance even further out of reach in just these circumstances.

III. A Simple Model

We construct a simple model that enables us explicitly to determine which investors will make use of the equity market and which of the debt market, and that enables us to calculate the effective marginal cost of capital. Because we wish to focus on the equity market, we assume bankers can perfectly discriminate among borrowers—indeed, the function of banks is to differentiate potential borrowers into their appropriate risk classes—but that the equity market treats all those seeking equity the same. (Thus, while Stiglitz-Weiss, 1981, were concerned with imperfect information in the credit market, we are concerned here with imperfect information in the equity market. In a sequel, we investigate a more general model incorporating both.) We make the following assumptions:

ASSUMPTION 1: *Each firm is characterized by a net cash flow, θ, from existing operations and a set of new investment opportunities whose return is $\varepsilon Q(K)$, where ε is a random variable, $E(\varepsilon)=1$, $var(\varepsilon)=\sigma_\varepsilon^2$ and K is the level of investment.*[2]

For expositional reasons, although firms are assumed to have different levels of θ, $Q(\cdot)$ is assumed to be the same for all firms. The parameter θ describes the "quality" or "value" of a particular firm and has a distribution $N(\theta)$ across firms.[3]

At the beginning of the period, firms announce their equity sales intentions, and V_0, each firm's market value, adjusts accordingly. Firms then sell (or do not sell) equity, determine the level investment, and finance any uncovered balance with debt. At the end of the period, the results of new investment are determined, some firms go bankrupt, and the values of θ are revealed for the remaining firms. The terminal value of each firm's equity is determined based on its observed value of θ. Managers' compensation depends on current market value and the share of terminal market value held by original shareholders, if the firm does not go bankrupt. In the event of bankruptcy, managers bear a known fixed cost.

Assuming risk neutrality, the firm acts as if its maximizes,

$$(1)\quad T=mV_0+(1-m)(V_0/(V_0+e)) \times(\theta+Q(K)-b(1+R))-cP_B,$$

where $b=K-e\equiv$ level of new borrowing, $R\equiv$ expected return on debt, $c\equiv$ cost which "bankruptcy" imposes on a firm's managers, $P_B\equiv$ probability of bankruptcy, and $m\equiv$ factor describing the weight that firms place on their initial as opposed to their terminal market value.

ASSUMPTION 2: *Bankruptcy occurs if*

$$(2)\qquad \theta+\varepsilon Q(K)<(1+R_0)b,$$

[2] For simplicity, existing net cash flows are assumed to be certain. Making existing cash flows uncertain would merely complicate the analysis and reinforce the basic results.

[3] The model as presented involves only a single period, but can be easily extended to a sequence of periods with independent θ draws in each period.

where $R_0 \equiv$ contractual rate of interest on a firm's debt $> R$.

ASSUMPTION 3: *Lenders are fully informed, risk neutral, and require an expected return R,*

$$(3)\quad (1+R)=(1+R_0)(1-P_B) + \int_0^{\varepsilon_0}[(\theta+\varepsilon Q(K))/b]\,dF(\varepsilon),$$

where $\varepsilon_0 = [(1+R_0)b-\theta]/Q(K) \equiv$ the value of ε below which bankruptcy occurs.

ASSUMPTION 4: *Equity investors are risk neutral and require an expected return R. They observe only the level of a firm's equity sales in determining V_0. Firms selling equity sell a common dollar amount e_0.*

The information structure of the model may appear restrictive, but is in fact quite general. Allowing equity investors to observe only the level of equity sales is a matter of interpreting the model as applying to a set of firms whose other observable characteristics are identical. The analysis need only be replicated for each such class of firms to cover the full firm population.[4,5]

A firm's equity sale decision rule can be characterized by examining the function,

$$H(\theta)\equiv T^D(\theta)-T^E(\theta),$$

where

$$T^D \equiv mV_0^D+(1-m)$$
$$\times\big(\theta+Q(K^D)-K^D(1+R)\big)-cP_B^D,$$

$V_0^D \equiv$ initial value of firms selling no equity, $K^D \equiv$ optimal level of investment for a nonequity selling firm of quality θ (the θ argument has been supressed), $P_B^D \equiv$ level of bankruptcy risk implied by the optimal investment decisions of a nonequity issuing firm (again the θ argument has been suppressed), and

$$T^E \equiv mV_0^E+(1-m)\big(V_0^E/(V_0^E+e_0)\big)$$
$$\big(\theta+Q(K^E)-(K^E-e_0)(1+R)\big)-cP_B^E,$$

where V_0^E, K^E, and P_B^E are defined analogously to V_0^D, K^D, and P_B^D. Assuming that ε_0 lies in the lower tail of a single peeked ε distribution, it is relatively straightforward to show that $dH(\theta)/d\theta>0$. Thus, the optimal decision rule for individual firms on equity sales policy is the following,

$$(4)\qquad e^*=\begin{cases} e_0 & \text{if } \theta \le \hat\theta \\ 0 & \text{if } \theta>\hat\theta,\end{cases}$$

where $\hat\theta$ is defined by $H(\hat\theta)=0$. Given equation (4), *firms entering the equity market will be adversely selected.* And, although in this simple model an equilibrium always exists, it may be one with zero equity sales. However, if $\theta \ge 0$ for all firms and m is close to zero, then an equilibrium with positive equity sales will exist. In such an equilibrium, V_0^E is determined by the equation,

$$V_0^E=1/N_E\int_0^{\hat\theta}\big((\theta+Q(K^E))/(1+R)$$
$$-(K^E-e_0)\big)N(\theta)\,d\theta-e_0,$$

where

$$N_E=\int_0^{\hat\theta}N(\theta)\,d\theta.$$

It is relatively easy to show that the resulting equilibrium level of V_0^E has the following properties, under suitable regularity conditions on F,

(i) $dV_0^E/dc \ge 0$, (ii) $dV_0^E/dm \le 0$,

(iii) $d(HR)V_0^E/dR \ge 0$, (iv) $dV_0^E/d\sigma_\varepsilon^2 \ge 0$.

[4] The restriction to discrete levels of equity sales, though made primarily for expositional convenience, has certain important theoretical justifications and consequences.

[5] In order that each class include firms with more than a single value of θ, neither K nor b may be perfectly observable to equity investors. However, given current accounting conventions and the timing of debt reports, this is not implausible.

In each instance an increase in V_0^E is associated with an increase in the number of firms issuing equity (a decrease in V_0^E is associated with a decrease in the number of firms issuing equity).

The optimal investment condition which characterizes nonequity issuing firms is[6]

$$(5)\quad Q_k = (1+R)\left[1+(c/(1-m))\left(1/(f_0^D-P_B^D)Q\right) \times \left(1-(K\cdot Q_k/Q)(1-\theta/K(1+R))\right)\right],$$

where f_0^D is the level of the ε density function at ε_0^D and ε_0^D is the level of the ε for a nonequity issuing firm at which, when K is optimally chosen, the firm defaults (ε_0^D depends of course on θ). The second bracketed term on the right-hand side of (5) represents the component of the cost of capital attributable to the marginal increase in the risk of bankruptcy. As θ falls (because a negative demand shock reduces the value of existing cash flows), this term may rise dramatically as f_0^D and P_B^D increase. Any such increase is limited ultimately by the possibility of issuing equity.

In practice, the "effective" cost of issuing equity may be so high as to be prohibitive, because any firm that issues equity obtains a "bad" label. Recent studies (for example, Paul Asquith and David Mullins, 1983) indicate that an equity issue announcement reduces the value of a firm by about 3 percent. And this may be a substantial underestimate since it is based on firms who actually issue equity and who are, as a result, likely to have the lowest cost of doing so. Thus, if $m = 1/2$ and a new equity issue amounts to 5 percent of a firm's outstanding stock, the signalling cost of equity will, by itself, amount to more than 30 percent. It is not surprising, therefore, that firms rarely issue equity. Moreover, if strong firms enjoy an enhanced advantage over weak ones in the face of adverse economic conditions, a negative economic surprise will increase the dispersion of $N(\theta)$ and increase the cost of issuing equity just when it is most needed.[7]

IV. Concluding Remarks

Informational imperfections have a fundamental effect on the functioning of the capital market. In some circumstances, competitive markets will be characterized by credit rationing: it is the availability of capital and not its cost that determines the level of investment. Here we have provided an explanation for why firms whose credit is constrained do not avail themselves of the equity market. We have also shown that the effective marginal cost of capital for those who are not constrained is not simply related either to the real long-term interest rate (as Keynes' hypothesized), or to the price of equity (as more recent portfolio theories have argued); the effective marginal cost of capital may experience much larger cyclical fluctuations than either of these variables. These variations in the effective cost of capital in turn play an important role in explaining observed patterns of cyclical behavior regarding both investment and prices.

Although the former effect is obvious, the latter may not be. When current prices affect not only present but future demand (see Edmund Phelps and Sidney Winter, 1970), firms will maximize profits with a price at which short-run marginal costs lie above short-run marginal revenues. The gap is filled by the contribution of lower prices to future profits. Under these circumstances, an increase in the cost of capital reduces the present value of any future market position and will lead to an increase in current prices. Our cost of capital view leads to just such a conclusion; as a recession begins, this tendency toward higher prices might well counteract the effect of falling demand and account for some price stickiness. In this and other ways, informational imperfections may provide a consistent economic explanation

[6] A similar condition would apply for equity issuing firms because they are limited to issuing only e_0 dollars of equity. This is an artifact of our assumptions.

[7] Under suitable regularity conditions on $N(\theta)$, an increase in the dispersion of $N(\theta)$ reduces V_0^F and increases the cost of issuing equity.

for many hitherto unexplained aspects of macroeconomic behavior.

REFERENCES

Asquith, Paul and Mullins, David W., "Equity Issues and Stock Price Dilution," unpublished paper, Harvard Business School, November 1983.

Phelps, Edmund S. and Winter, Sidney, G., "Optimal Price Policy under Atomistic Competition" in Edmund S. Phelps, ed., *Microeconomic Foundations of Employment and Inflation Theory*, New York: W. W. Norton and Co., 1970.

Ross, Stephen A., "The Determination of Financial Structure: The Incentive Signalling Approach," *Bell Journal of Economics*, Spring 1977, *8*, 23–40.

Stiglitz, Joseph E. and Weiss, Andrew, "Credit Rationing and Markets with Imperfect Information," *American Economic Review*, June 1981, *71*, 393–411.

______ **and** ______, "Incentive Effects of Terminations: Applications to the Credit and Labor Markets," *American Economic Review*, December 1983, *73*, 912–27.

Efficiency Wage Models of Unemployment

By JANET L. YELLEN*

Keynesian economists hold it to be self-evident that business cycles are characterized by involuntary unemployment. But construction of a model of the cycle with involuntary unemployment faces the obvious difficulty of explaining why the labor market does not clear. Involuntarily unemployed people, by definition, want to work at less than the going wage rate. Why don't firms cut wages, thereby increasing profits?

This paper surveys a recent literature which offers a convincing and coherent explanation why firms may find it unprofitable to cut wages in the presence of involuntary unemployment. The models surveyed are variants of the efficiency wage hypothesis, according to which, labor productivity depends on the real wage paid by the firm. If wage cuts harm productivity, then cutting wages may end up raising labor costs. Section I describes some of the general implications of the efficiency-wage hypothesis in its simplest form. Section II describes four distinct microeconomic approaches which justify the relation between wages and productivity. These approaches identify four benefits of higher wage payments: reduced shirking by employees due to a higher cost of job loss; lower turnover; an improvement in the average quality of job applicants; and improved morale.[1] Section III explains how the efficiency-wage hypothesis, with near rational behavior, can explain cyclical fluctuations in unemployment.

I. The Efficiency Wage Hypothesis

The potential relevance of the efficiency-wage hypothesis in explaining involuntary unemployment and other stylized labor market facts can be seen in a rudimentary model. Consider an economy with identical, perfectly competitive firms, each firm having a production function of the form $Q = F(e(\omega)N)$, where N is the number of employees, e is effort per worker, and ω is the real wage. A profit-maximizing firm which can hire all the labor it wants at the wage it chooses to offer (see Joseph Stiglitz, 1976a; Robert Solow, 1979), will offer a real wage, ω^*, which satisfies the condition that the elasticity of effort with respect to the wage is unity. The wage ω^* is known as the efficiency wage and this wage choice minimizes labor cost per efficiency unit. Each firm should then optimally hire labor up to the point where its marginal product, $e(\omega^*)F'(e(\omega^*)N^*)$, is equal to the real wage, ω^*. As long as the aggregate demand for labor falls short of aggregate labor supply and ω^* exceeds labor's reservation wage, the firm will be unconstrained by labor market conditions in pursuing its optimal policy so that equilibrium will be characterized by involuntary unemployment. Unemployed workers would strictly prefer to work at the real wage ω^* than to be unemployed, but firms will not hire them at that wage or at a lower wage. Why? For the simple reason that any reduction in the wage paid would lower the productivity of all employees already on the job. Thus the efficiency-wage hypothesis explains involuntary unemployment.

Extended in simple ways this hypothesis also explains four other labor market phenomena: real wage rigidity; the dual labor market; the existence of wage distributions for workers of identical characteristics; and discrimination among observationally distinct groups. Concerning real wage rigidity, in the simple model just described, real shocks which shift the marginal product of labor alter employment, but not the real wage. In more elaborate versions of the model discussed below, such shocks will change the real wage, but not sufficiently to leave unemployment unaltered.

*University of California, Berkeley, CA 94720. I am indebted to George Akerlof, David Estenson, Michael Reich, and James Wilcox for invaluable discussion and comments.

[1] For a previous survey of portions of this literature, see Guillermo Calvo (1979).

Dual labor markets can be explained by the assumption that the wage-productivity nexus is important in some sectors of the economy, but not in others. For the primary sector, where the efficiency-wage hypothesis is relevant, we find job rationing and voluntary payment by firms of wages in excess of market clearing; in the secondary sector, where the wage-productivity relationship is weak or nonexistent, we should observe fully neoclassical behavior. The market for secondary-sector jobs clears, and anyone can obtain a job in this sector, albeit at lower pay. The existence of the secondary sector does not, however, eliminate involuntary unemployment (see Robert Hall, 1975), because the wage differential between primary- and secondary-sector jobs will induce unemployment among job seekers who choose to wait for primary-sector job openings.

Theorists who emphasize the importance of unemployment due to the frictions of the search process have frequently found it difficult to explain the reasons for a distribution of wage offers in the market. The efficiency-wage hypothesis also offers a simple explanation for the existence of wage differentials which might motivate the search process emphasized by Edmund Phelps and others. If the relationship between wages and effort differs among firms, each firm's efficiency wage will differ, and, in equilibrium, there will emerge a distribution of wage offers for workers of identical characteristics.

The efficiency-wage hypothesis also explains discrimination among workers with different observable characteristics. This occurs if employers simply prefer, say, men to women. With job rationing, the employer can indulge his taste for discrimination at zero cost. As another possibility, employers may know that the functions relating effort to wages differ across groups. Then each group has its own efficiency wage and corresponding "efficiency labor cost." If these labor costs differ, it will pay firms to hire first only employees from the lowest cost group. Any unemployment that exists will be confined to labor force groups with higher costs per efficiency unit. With fluctuations in demand, these groups will bear a disproportionate burden of layoffs.

II. Microfoundations of the Efficiency-Wage Model

Why should labor productivity depend on the real wage paid by firms? In the *LDC* context, for which the hypothesis was first advanced, the link between wages, nutrition, and illness was emphasized. Recent theoretical work has advanced a convincing case for the relevance of this hypothesis to developed economies. In this section, four different microeconomic foundations for the efficiency-wage model are described and evaluated.

A. *The Shirking Model*

In most jobs, workers have some discretion concerning their performance. Rarely can employment contracts rigidly specify all aspects of a worker's performance. Piece rates are often impracticable because monitoring is too costly or too inaccurate. Piece rates may also be nonviable because the measurements on which they are based are unverifiable by workers, creating a moral hazard problem. Under these circumstances, the payment of a wage in excess of market clearing may be an effective way for firms to provide workers with the incentive to work rather than shirk. (See Samuel Bowles, 1981, 1983; Guillermo Calvo, 1979; B. Curtis Eaton and William White, 1982; Herbert Gintis and Tsuneo Ishikawa, 1983; Hajime Miyazaki, forthcoming; Carl Shapiro and Stiglitz, 1982; and Steven Stoft, 1982.) The details of the models differ somewhat, depending on what is assumed measurable, at what cost, and the feasible payment schedules.

Bowles, Calvo, Eaton-White, Shapiro-Stiglitz, and Stoft assume that it is possible to monitor individual performance on the job, albeit imperfectly. In the simplest model, due to Shapiro-Stiglitz, workers can decide whether to work or to shirk. Workers who shirk have some chance of getting caught, with the penalty of being fired. This has been termed "cheat-threat" theory by Stoft because, if there is a cost to being fired, the threat of being sacked if caught cheating creates an incentive not to shirk. Equilibrium then entails unemployment. If all firms pay an identical wage, and if there is full employment, there would be no cost to shirking and it would pay all workers, assumed to get pleasure from loafing on the job, to shirk. In these circumstances, it pays each firm to raise its wage to eliminate shirking. When all firms do this, average wages rise and employment falls. In equilibrium, all firms pay the same wage above market clearing, and unemployment, which makes job loss costly, serves as a worker-discipline device. Unemployed workers cannot bid for jobs by offering to work at lower wages. If the firm were to hire a worker at a lower wage, it would be in the worker's interest to shirk on the job. The firm knows this and the worker has no credible way of promising to work if he is hired.

The shirking model does *not* predict, counterfactually, that the bulk of those unemployed at any time are those who were fired for shirking. If the threat associated with being fired is effective, little or no shirking and sacking will actually occur. Instead, the unemployed are a rotating pool of individuals who have quit jobs for personal reasons, who are new entrants to the labor market, or who have been laid off by firms with declines in demand. Pareto optimality, with costly monitoring, will entail some unemployment, since unemployment plays a socially valuable role in creating work incentives. But the equilibrium unemployment rate will not be Pareto optimal (see Shapiro-Stiglitz).

In contrast to the simple efficiency-wage model, the shirking model adds new arguments to the firm's effort function—the average wage, aggregate unemployment, and the unemployment benefit. The presence of the unemployment rate in the effort function yields a mechanism whereby changes in labor supply affect equilibrium wages and employment. New workers increase unemployment, raising the penalty associated with being fired and inducing higher effort at any given wage. Firms accordingly lower wages and hire more labor as a result. In a provocative recent paper, Thomas Weisskopf, Bowles, and David Gordon (1984) have used the presence of the unemployment benefit in the effort function to explain the secular decline in productivity in the United States; they argue that a major part of the productivity slowdown is attributable to loss of employer control due to a reduction in the cost of job loss. The shirking model also offers an interpretation of hierarchical wage differentials, in excess of productivity differences (Calvo and Stanislaw Wellisz, 1979).

All these models suffer from a similar theoretical difficulty—that employment contracts more ingenious than the simple wage schemes considered, can reduce or eliminate involuntary unemployment. In the cheat-threat model, the introduction of employment fees allows the market to clear efficiently as long as workers have sufficient capital to pay them (see Eaton-White and Stoft). Unemployed workers would be willing to pay a fee to gain employment. Fees lower labor costs, giving firms an incentive to hire more workers. If all firms charge fees, any worker who shirks and is caught knows that he will have to pay another fee to regain employment. This possibility substitutes for the threat of unemployment in creating work incentives. Devices which function similarly are bonds posted by workers when initially hired and forfeited if found cheating, and fines levied on workers caught shirking. The threat of forfeiting the bond or paying the fine substitutes for the threat of being fired. Edward Lazear (1981) has demonstrated the use of seniority wages to solve the incentive problem. Workers can be paid a wage less than their marginal productivity when they are first hired with a promise that their earnings will later exceed their marginal productivity. The upward tilt in the age-earnings profile provides a penalty for shirking; the present value of the wage paid can fall to the market-clearing level, eliminating involuntary unemployment.

As a theoretical objection to these schemes, employers would be subject to moral hazard in evaluating workers' effort. Firms would have an obvious incentive to declare workers shirking and appropriate their bonds, collect fines, or replace them with new fee-paying workers. In Lazear's model, in which the firm pays a wage in excess of marginal product to senior workers, there is an incentive for the firm to fire such workers, replacing them with young workers, paid less than their productivity. The seriousness of this moral hazard problem depends on the ability of workers to enforce honesty on the firm's part. If effort is observable both by the firm and by the worker, and if it can be verified by outside auditors, the firm will be unable to cheat workers. Even without outside verification, Lazear has shown how the firm's concern for its reputation can overcome the moral hazard problem. Sudipto Bhattacharya (1983) has suggested tournament contracts that also overcome the moral hazard problem. The firm can commit itself to a fixed wage plan in which a high wage is paid to a fraction of workers and a low wage to the remaining fraction according to an *ex post*, possibly random, ranking of their effort levels. By precommitting itself to such a plan with a fixed wage bill, any moral hazard problem on the firm's part disappears.

B. *The Labor Turnover Model*

Firms may also offer wages in excess of market clearing to reduce costly labor turnover. (See Steven Salop, 1979; Ekkehart Schlicht, 1978; and Stiglitz, 1974.) The formal structure of the labor turnover model is identical to that of the shirking model. Workers will be more reluctant to quit the higher the relative wage paid by the current firm, and the higher the aggregate unemployment rate. If all firms are identical, one possible equilibrium has all firms paying a common wage above market clearing with involuntary unemployment serving to diminish turnover.

The theoretical objection to the prediction of involuntary unemployment in this model again concerns the potential for more sophisticated employment contracts to provide Pareto-superior solutions. As Salop explains, the market for new hires fails to clear because an identical wage is paid to both trained and untrained workers. Instead, new workers could be paid a wage equal to the difference between their marginal product and their training cost. A seniority wage scheme might accomplish this, although, if training costs are large and occur quickly it might prove necessary to charge a fee to new workers. In contrast to the shirking model, an employment or training fee scheme could be employed without the problem of moral hazard. It is no longer in any firm's interest to dismiss trained workers; explicit contracts could probably be written to insure that training is actually provided to fee paying workers. Although moral hazard thus appears to be a less formidable barrier to achieving neoclassical outcomes via fees or bonds than in the shirking model, capital market imperfections or institutional or sociological constraints may in fact make them impractical.

C. *Adverse Selection*

Adverse selection yields further reason for a relation between productivity and wages. Suppose that performance on the job depends on "ability" and that workers are heterogeneous in ability. If ability and workers' reservation wages are positively correlated, firms with higher wages will attract more able job candidates. (See James Malcolmson, 1981; Stiglitz, 1976b; Andrew Weiss, 1980.) In such a model, each firm pays an efficiency wage and optimally turns away applicants offering to work for less than that wage. The willingness of an individual to work for less than the going wage places an upper bound on his ability, raising the firm's estimate that he is a lemon. The model provides an explanation of wage differentials and different layoff probabilities for observationally distinct groups due to statistical discrimination if it is known that different groups have even slight differences in the joint distributions of ability and acceptance wages. However, for the adverse-selection model to provide a convincing account of involuntary unemployment, firms must be unable to measure effort and pay piece rates after workers are hired, or to fire workers whose output is too low. Clever firms may also be able to mitigate adverse selection in hiring by designing self-selection or screening devices which induce workers to reveal their true characteristics.

D. *Sociological Models*

The theories reviewed above are neoclassical in their assumption of individualistic maximization by all agents. Solow (1980) has argued, however, that wage rigidity may more plausibly be due to social conventions and principles of appropriate behavior that are not entirely individualistic in origin. George Akerlof (1982) has provided the first explicitly sociological model leading to the efficiency-wage hypothesis. He uses a variety of interesting evidence from sociological studies to argue that each worker's effort depends on the work norms of his group. In Akerlof's partial gift exchange model, the firm can succeed in raising group work norms and average effort by paying workers a gift of wages in excess of the minimum required, in return for their gift of effort above the minimum required. The sociological model can explain phenomena which seem inexplicable in neoclassical terms—why firms don't fire workers who turn out to be less productive, why piece rates are avoided even when feasible, and why firms set work standards exceeded by most workers. Akerlof's paper in this issue explores alternative sociological foundations for the efficiency wage hypothesis. Sociological considerations governing the effort decisions of workers are also emphasized in Marxian discussions of the extraction of labor from labor power (see, for example, Bowles, 1983).

III. Explaining the Business Cycle

Any model of the business cycle must explain why changes in aggregate demand cause changes in aggregate employment and output. A potential problem of the efficiency-wage hypothesis in this regard is the absence of a link between aggregate demand and economic activity. In an economy with efficiency-wage setting, there is a positive natural rate of unemployment and real wage rigidity. But the economy's aggregate output is independent of price at this natural rate. These models have no wage or price stickiness to cause real consequences from aggregate demand shocks. However, for a natural but subtle reason, the efficiency-wage model is consistent with nominal wage rigidity and cyclical unemployment. This reason (suggested by Stoft), is explored in depth by Akerlof and myself (1983), where we argue that sticky wage and price behavior, that will cause significant business cycle fluctuations, is consistent with near rationality in an economy with efficiency wage setting. Any firm that normally chooses its wage as part of an optimizing decision will incur losses that are only second-order if it follows a rule of thumb in adjusting nominal wages which leads to a real wage error. At the point of maximum profits, the profit function relating wages to profits is flat. Thus, in the neighborhood of the optimum wage, the loss from wage errors is second-order small. This implies that firms with sticky wages have profits that are insignificantly different from firms with maximizing behavior. Furthermore, if firms have price-setting power because of downward-sloping demand curves, for similar reasons, price-setting errors also lead to insignificant losses.

In the Akerlof-Yellen model, firms are efficiency-wage setters and monopolistic competitors. In the long run, wages and prices are set by all firms in an optimal way. In the short run, in response to aggregate demand shocks, some firms keep nominal wages and prices constant, while other firms choose these variables optimally. In this model, a cut in the money supply causes a first-order change in employment, output, and profits. But the behavior of nonmaximizers is near rational in the sense that the potential gain any individual firm could experience by abandoning rule of thumb behavior is second-order small. And thus the efficiency-wage hypothesis can be extended into a full-fledged Keynesian model of the business cycle generated by sticky prices and wages.

IV. Concluding Remarks

It has been widely observed that the existence of excess labor supply does not lead to aggressive wage cutting by workers and firms. Firms appear content to pay workers more than the wages required by their potential replacements. The models surveyed here offer several different and plausible explanations of this seemingly paradoxical fact. In addition to accounting for the persistence of involuntary unemployment in competitive markets, these efficiency wage models can explain why unemployment varies in response to aggregate demand shocks. In sum, these models provide a new, consistent, and plausible microfoundation for a Keynesian model of the cycle.

REFERENCES

Akerlof, George, "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics*, November 1982, *97*, 543–69.

______, "Gift Exchange and Efficiency Wage Theory: Four Views," *American Economic Review Proceedings*, May 1984, *74*, 79–83.

______ **and Yellen, Janet,** "The Macroeconomic Consequences of Near Rational, Rule of Thumb Behavior," mimeo., University of California-Berkeley, September 1983.

Bhattacharya, Sudipto, "Tournaments and Incentives: Heterogeneity and Essentiality," mimeo., Stanford University, March 1983.

Bowles, Samuel, "Competitive Wage Determination and Involuntary Unemployment: A Conflict Model," mimeo., University of Massachusetts, May 1981.

______, "The Production Process in a Competitive Economy: Walrasian, Neo-Hobbesian and Marxian Models," mimeo., University of Massachusetts, May 1983.

Calvo, Guillermo, "Quasi-Walrasian Theories of Unemployment," *American Economic Review Proceedings*, May 1979, *69*, 102–07.

______ **and Wellisz, Stanislaw,** "Hierarchy, Ability and Income Distribution," *Journal of Political Economy*, October 1979, *87*, 991–1010.

Eaton, B. Curtis and White, William, "Agent Compensation and the Limits of Bonding," *Economic Inquiry*, July 1982, *20*, 330–43.

Gintis, Herbert and Ishikawa, Tsuneo, "Wages, Work Discipline and Macroeconomic Equilibrium," mimeo., 1983.

Hall, Robert, "The Rigidity of Wages and the Persistence of Unemployment," *Brookings Papers on Economic Activity*, 2:1975, 301–35.

Lazear, Edward, "Agency, Earnings Profiles, Productivity, and Hours Restrictions," *American Economic Review*, September 1981, *71*, 606–20.

Malcolmson, James, "Unemployment and the Efficiency Wage Hypothesis," *Economic Journal*, December 1981, *91*, 848–66.

Miyazaki, Hajime, "Work Norms and Involuntary Unemployment," *Quarterly Journal of Economics*, forthcoming.

Salop, Steven, "A Model of the Natural Rate of Unemployment," *American Economic Review*, March 1979, *69*, 117–25.

Schlicht, Ekkehart, "Labour Turnover, Wage Structure and Natural Unemployment," *Zeitschrift für die Gesamte Staatswissenschaft*, June 1978, *134*, 337–46.

Shapiro, Carl and Stiglitz, Joseph, "Equilibrium Unemployment as a Worker Discipline Device," mimeo., Princeton University, April 1982.

Solow, Robert, "Another Possible Source of Wage Stickiness," *Journal of Macroeconomics*, Winter 1979, *1*, 79–82.

______, "On Theories of Unemployment," *American Economic Review*, March 1980, *70*, 1–11.

Stiglitz, Joseph, "Wage Determination and Unemployment in L.D.C.'s: The Labor Turnover Model," *Quarterly Journal of Economics*, May 1974, *88*, 194–227.

______, (1976a) "The Efficiency Wage Hypothesis, Surplus Labour, and the Distribution of Income in L.D.C.s," *Oxford Economic Papers*, July 1976, *28*, 185–207.

______, (1976b) "Prices and Queues as Screening Devices in Competitive Markets," IMSSS Technical Report No. 212, Stanford University, August 1976.

Stoft, Steven, "Cheat-Threat Theory: An Explanation of Involuntary Unemployment," mimeo., Boston University, May 1982.

Weiss, Andrew, "Job Queues and Layoffs in Labor Markets with Flexible Wages," *Journal of Political Economy*, June 1980, *88*, 526–38.

Weisskopf, Thomas, Bowles, Samuel and Gordon, David, "Hearts and Minds: A Social Model of Aggregate Productivity Growth in the U.S., 1948–1979," *Brookings Papers on Economic Activity*, 1984, forthcoming.

Recent Changes in Macro Policy and its Effects: Some Time-Series Evidence

By JOHN B. TAYLOR*

Has the macroeconomic policy "regime" changed in the United States in the last few years? If so, has this change had an effect on macroeconomic relationships, and, in particular, on the relationship between inflation and unemployment? These two questions have been on the minds of most macroeconomists since the October 1979 switch from interest rates to reserves as an intermediate target for the Federal Reserve and the explicit endorsement of monetarism by the Reagan Administration in early 1981. The questions are important not only for predicting the course of the economy in the future, but also for assessing the practical significance of the Lucas critique.[1] According to this critique a change in policy regime should cause changes in the way the economy operates. Comparing economic relations before and after a change in regime should provide a test of the critique.

The approach in many recent studies has been to assume that the answer to the first question is yes—that there has been a significant change in the macroeconomic policy regime—and to look for changes in the Phillips curve as a test of the Lucas critique. Thus far, the results have been mixed. Otto Eckstein (1983), Steven Englander and Cornelis Los (1983), and George Perry (1983) conclude that there has been no significant change in the Phillips curve relationship. Phillip Cagan and William Fellner (1983) and Wayne Vroman (1983) find some evidence of a change: wage inflation has come down more quickly—especially in 1982 and early 1983—than would be predicted from a Phillips curve.

However, the assumption that there has been a regime change has received little empirical attention. The October 1979 switch from interest rate targetting to reserve targetting at the Fed is usually taken as prima facie evidence of a significant change in policy regime. But such changes in operating procedures per se do not necessarily entail a change in policy regime relevant for macroeconomic purposes. If they did, then the renewed emphasis on interest rates rather than reserves starting in late 1982 should be cited as evidence of a return to the old regime—a view which few researchers have taken. In fact, studies have shown that either interest rates or reserves can be used as intermediate targets for controlling the money supply and ultimately aggregate demand. (The choice between the two depends on whether shocks to the money markets are from the demand side or the supply side.) The issue of importance for assessing whether a regime change has taken place is how much the Fed reacts by adjusting the money supply (appropriately defined) in response to conditions in the economy. The procedure it uses to bring about this response is irrelevant.[2] Implicit in a macroeconomic policy regime is a "rule" relating the money supply to economic con-

*404 Woodrow Wilson School, Princeton University, Princeton, NJ 08544. This research was supported by a grant from the National Science Foundation at the National Bureau of Economic Research and conducted in part at the Federal Reserve Bank of Philadelphia. I am grateful to Steve Fries for research assistance, and to Steven Englander and Cornelis Los for comments on an earlier draft.

[1] See Robert Lucas (1976). William Fellner's (1978) credibility hypothesis is similar to the Lucas Critique in this context and has been tested in similar ways.

[2] Some have argued that, for political reasons, by focusing on reserves the Fed would be able to let interest rates go higher than under interest rate targetting. With reserve targetting, Fed officials could shift the blame for high interest rates elsewhere.

ditions. The fact that the last three years are a suitable time for testing the Lucas Critique can be established only by providing evidence that this rule has changed significantly.

The aim of this paper is to use simple time-series techniques to examine whether there has been a significant change in macroeconomic policy in the special sense described above, and whether, as a result of this change, the relationship between inflation and unemployment has shifted in ways suggested by recent theories of wage and price determination. To do this a small vector autoregression is estimated and transformed into a pure moving average form in which changes in macro policy and its effects can be measured.

I. Measuring The Change in Aggregate Demand Policy

Financial innovations such as sweep accounts and the move toward deregulation of financial services have made it difficult to use any one measure of the money supply for estimating a policy rule in recent years. The relationship between money and nominal *GNP* has been even more volatile than usual as a result of these innovations and regulatory changes. An alternative aproach is needed.

The most direct way to get around the velocity shift problem is to focus on aggregate demand itself as measured by nominal *GNP*. A change in policy regime would then occur if the Fed altered its rule for adjusting its nominal *GNP* targets. In particular, the Fed would become less *accommodative* in the 1980's if it did not permit nominal *GNP* to increase as much in response to inflation shocks as it has in the past. Such a less-accommodative policy would permit real *GNP* to fall by a larger amount in response to inflation shocks. Alternatively stated, it would permit unemployment to rise by a larger amount in response to inflation shocks.

During the late 1960's and 1970's the Fed allowed the growth rate of nominal *GNP* to *rise* to about 12 percent, which accommodates a steady 9 percent inflation if potential *GNP* grows at 3 percent. If there has been a change in the policy regime it is likely to be to one in which the Fed will aim to make nominal *GNP* less accommodative to inflation. Perhaps it will eventually aim to keep the growth rate of nominal *GNP constant* at a lower level (say 5 percent) as some economists have recently proposed. These less-accommodative policies would entail larger increases in unemployment in response to inflation shocks.

A reaction function for measuring this change can be obtained by regressing the unemployment rate on past inflation rates and past unemployment rates. Adding lagged unemployment terms to the equation allows for reaction lags or other reasons for delayed response by the Fed. If the coefficients on the past inflation rates in this equation have increased in the last few years, then this would be evidence of a change toward a less accommodative policy regime.

II. Measuring Change in the Inflation Process

How would the inflation process change as a result of such a change to a less-accommodative policy? Recent research on wage and price determination (see my 1980 article for example) suggests that inflation will be less persistent (i.e., have smaller positive autocorrelation) under a less-accommodative aggregate demand policy. A shock to the inflation rate will not have as prolonged an effect on inflation with a less-accommodative policy. This research also suggests that the effect of a given level of unemployment on inflation will increase. This is due to the expectation that the Fed will not accommodate inflation in the future. Expectations of higher unemployment in the future, as well as expectations of lower inflation in the future, will tend to reduce inflation today.

These potential changes can be measured by regressing the inflation rate on past inflation rates and on past unemployment rates. If the coefficients on past inflation rates decline and the coefficients on the unemployment rates increase in absolute value, while at the same time macro policy becomes less accommodative, then we would have evidence which is consistent with this theory of wage and price determination.

III. The Evidence from Vector Autoregressions

The regressions suggested above for measuring changes in macro policy and in the inflation process can be combined to form a simple vector autoregression in which the vector of unemployment and inflation observations is regressed on past values of itself. This bivariate autoregression was estimated over two sample periods: 1954:1–1981:1 and 1954:1–1983:3. The longer sample period includes the recent disinflation period where the effects of a regime change should be noticeable. Unfortunately the period 1981:1–1983:3 is too short to estimate a vector autoregression separately.

Let $y_t = (w_t, u_t)'$ where w_t is the rate of change in average hourly earnings index (a measure of inflation), and u_t is the unemployment rate for prime age males (a demographically stable measure of unemployment).[3] The estimated equations are of the form

$$y_t = \sum_{i=1}^{4} A_i y_{t-i} + \varepsilon_t, \tag{1}$$

and are based on quarterly observations. The moving average for this system is given by

$$y_t = \sum_{i=0}^{\infty} \theta_i \varepsilon_{t-i}, \tag{2}$$

where the θ_i matrices can be obtained directly from the A_i matrices.[4] As Christopher Sims (1980) has argued, the moving average coefficients usually show a smoother pattern than the autoregressive coefficients and are therefore easier to interpret.

The autoregressions for the two periods are reported in Table 1 and the moving average representation is shown in Figure 1. In assessing the change in macro policy and in the inflation process we can compare these two vector autoregressions and their moving average representations. This procedure is not entirely satisfactory since the sample period which contains the last eleven quarters also includes the earlier period; if there has been a change in the last eleven quarters, then these estimates will be a mixture of two dissimilar periods between which a structural change has occurred. Assuming that the new policy is maintained, future studies may be able to utilize additional observations to estimate separate autoregressions and moving average representations under the new regime.

TABLE 1—BIVARIATE AUTOREGRESSIONS FOR WAGE INFLATION AND UNEMPLOYMENT

	Lags	1	2	3	4	*SSR*
Sample: 1954:1–1981:1						
w-equation	*w*	.40	.26	.12	.22	
	u	.03	.02	.02	−.05	98.0
u-euqation	*w*	−.45	.56	−.51	.30	
	u	1.62	−.89	.21	−.01	11.1
Correlation of Residuals: $\rho_{wu} = -.290$						
Sample 1954:1–1983:3						
w-equation	*w*	.41	.26	.10	.19	
	u	.03	.01	.04	−.04	105.2
u-equation	*w*	−.50	.65	−.63	.28	
	u	1.66	−.88	.16	.03	13.4
Correlation of Residuals: $\rho_{wu} = -.306$						

Note: *w* is the percentage quarterly change in the average hourly earnings index measured at an annual rate, *u* is the unemployment rate for males 25 to 54; *SSR* is the sum of squared residuals for each equation; the equations were estimated with a constant term.

Visual inspection of the autoregressive equations in Table 1 indicates that the effect of adding additional sample points during the disinflation period has a relatively small effect on the autoregressive coefficients. The *F*-tests for structural homogeneity give values of .735 and 2.07 for the inflation and unemployment equations, respectively. The 5 percent significance point is 1.94. Hence there is more evidence of a structural change in the policy equation than in the behavioral equation for inflation. But neither change is dramatic.

Note that the changes in the coefficients do indicate a shift to a less-accommodative policy, and that the effects of this change are exactly as predicted above. The sum of the coefficients on lagged wages in the unem-

[3] The nominal wage series prior to 1961 was provided by Robert Gordon. See his 1971 paper.

[4] The covariance matrix of the ε_t in this representation has not been orthogonalized and is equal to the covariance matrix estimated in equation (1). Sims discusses alternative ways of presenting the moving average representation.

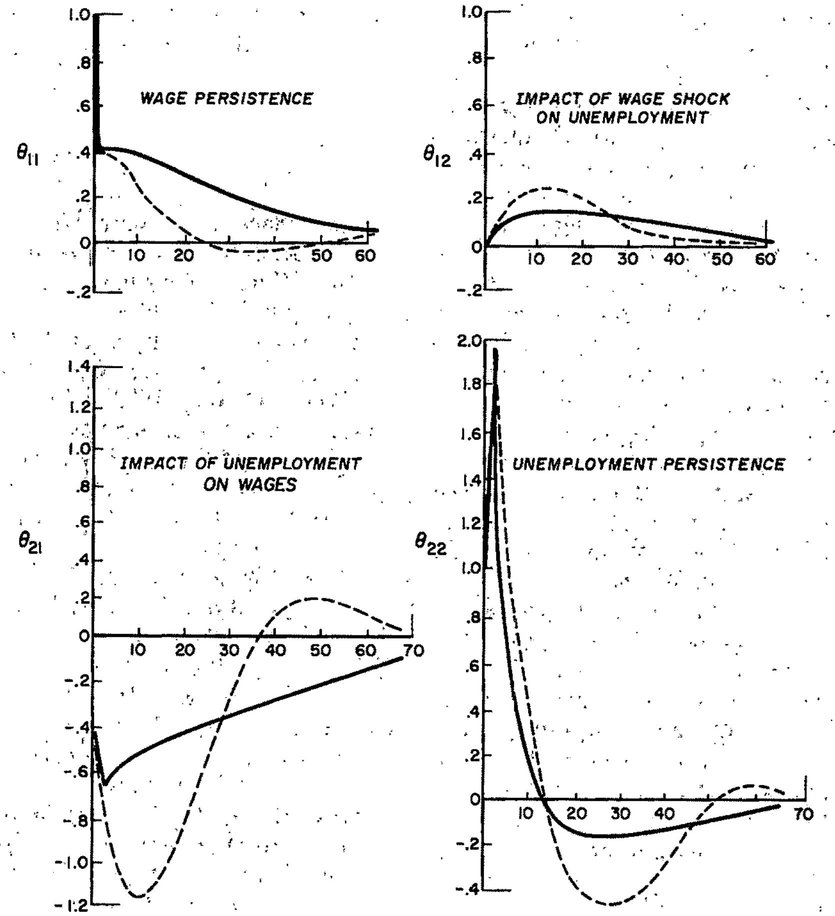


FIGURE 1. MOVING AVERAGE REPRESENTATIONS FOR WAGE GROWTH AND UNEMPLOYMENT

Note: 1954:1–1981:1 ——; 1954:1–1983:3 ---

ployment equation increases, the sum of the coefficients on lagged wages in the wage equation declines, and the sum of the coefficients on unemployment in the wage equation increases in absolute value.

It is likely that the lack of strong statistical significance in these equations is due to the small number of observations in the post-1980 period and the resulting low power of the structural homogeneity tests at conventional significance levels. If so, then if policy remains less accommodative in the future, the additional observations will strengthen the statistical significance. The striking fact that the direction of the changes correspond to what would be predicted from a less-accommodative policy is suggestive that this may be the case.

The moving average coefficients shown in the four charts in Figure 1 suggest the same interpretation as the simple comparison of the sum of lagged coefficients. They also provide information about how the dynamics of the economy have changed. The charts in Figure 1 are normalized so that zero is the steady-state value for each variable. For example, the normal level of the unemployment rate is zero.

The chart in the upper right-hand corner of the figure shows that the initial impact of a wage shock on unemployment is larger for the sample that includes the post-1980 data. The impact remains larger for 25 quarters, but eventually the response of unemployment falls below that in the earlier period. According to these estimates, policy seems to have become less accommodative, not only by responding by a larger amount to inflation shocks, but also by responding more quickly. This policy would generate recessions which are larger than under the old policy, but which do not last as long.

Examining the upper left-hand portion it is clear that wage inflation is much less persistent for the sample that includes the last three years. The impact of any shock to wages deteriorates more rapidly than in the pre-1980 sample. Finally the impact of unemployment on inflation is shown in the lower left-hand corner of the figure. The initial and medium-term impacts of unemployment on wage formation are larger for the sample which includes the last three years. Eventually, however, the effect of the unemployment shock is smaller.

IV. Concluding Remarks

The purpose of this paper has been to look for changes in the behavior of macroeconomic policy and the relation between inflation and unemployment during the last three years using time-series techniques. There is evidence that the policy "rule" has changed in the direction of a less-accommodative policy and that the inflation process has become less persistent. Under the less-accommodative policy, the increase in unemployment caused by inflation "shocks" is larger but less prolonged.

The results, however, are based on a relatively small number of observations during the new regime and are not strongly significant statistically. While the results are in general agreement with recent theories of the effect of macro policy on wage and price formation, they must therefore be viewed as tentative.

REFERENCES

Cagan, Phillip and Fellner, William J., "Tentative Lessons from the Recent Disinflationary Effort," *Brookings Papers on Economic Activity*, 2:1983, 603–08.

Eckstein, Otto, "Disinflation," *Alternatives for the 1980s*, No. 10, Center for National Policy, October 1983.

Englander, A. Steven and Los, Cornelis A., "The Stability of the Phillips Curve and its Implications for the 1980s," Research Paper No. 8303, Federal Reserve Bank of New York, February 1983.

Fellner, William J., "The Core of the Controversy about Reducing Inflation: An Introductory Analysis," in his *Contemporary Economic Problems*, Washington: American Enterprise, 1978.

Gordon, Robert J., "Inflation in Recession and Recovery," *Brookings Papers on Economic Activity*, 1:1971, 105–66.

Lucas, Robert E. Jr., "Econometric Policy Evaluation: A Critique," in Karl Brunner and Alan Meltzer, eds., *The Phillips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl. 1976, 19–46.

Perry, George L., "What Have We Learned about Disinflation?," *Brookings Papers on Economic Activity*, 2:1983, 587–602.

Sims, Christopher A., "Macroeconomics and Reality," *Econometrica*, January 1980, *48*, 1–48.

Taylor, John B., "Output and Price Stability: An International Comparison," *Journal of Economic Dynamics and Control*, February 1980, *2*, 109–32.

Vroman, Wayne, *Wage Inflation: Prospects for Deceleration*, Washington: The Urban Institute Press, 1983.

The Lucas Critique and the Volcker Deflation

By OLIVIER J. BLANCHARD*

Robert Lucas (1976) warned us that our econometric models were, by their very design, likely to perform poorly in the face of policy regime changes. The U.S. economy has in the last four years experienced precisely such a change, namely a change in monetary policy. Now is therefore a good time to study how two of the central macroeconometric equations, the Phillips curve and the term structure of interest rates, have fared during that period.

In this paper I present an informal account of the policy change and of its effects. I then discuss how we might expect the Phillips curve and the term structure equations to shift in the face of such a policy change. Empirical examinations of the behavior of pre-1979 Phillips curves and term structure equations in the last four years follow.

I. A Brief Description of Events

In October 1979, the Fed announced a change in monetary policy. Technically, the change was only in operating procedures, a shift from interest rate to money stock targeting; target growth ranges for the monetary aggregates were left unchanged.[1] This technical change was however intended to be both a signal of the Fed's commitment, and a prerequisite to lower money growth and inflation.

In retrospect, there is little doubt that this commitment was a serious one. Assessing the reality of the change was however much more difficult in the period following October 1979. Faced with what it perceived to be autonomous velocity shifts, the Fed chose partial accommodation, leading to large and erratic fluctuations in monetary aggregates.[2] As a result, the policy change was neither instantaneously perceived nor instantaneously believed. Direct evidence on the beliefs of financial markets suggests the following:[3] from October 1979 to December 1980, there was considerable doubt as to whether the Fed was committed to the reduction of inflation; in particular, doubts were fueled, in the spring and summer of 1980, by a decrease in short-term nominal rates in the face of the first recession. This decrease was interpreted by many as showing the unwillingness of the Fed to accept the recession. Doubts seem to have disappeared in the first half of 1981, due partly to the election of a new administration and partly to the Fed's policy of high interest rates. The increase in interest rates in June 1981, in the face of a second recession, seems to have been decisive in shifting financial markets' beliefs. Direct evidence on the beliefs of labor market participants is sparse; there is little to suggest that the change in policy was explicitly taken into account in labor contract negotiations, apart from its effects through unemployment.

This paper is not the place to give a thorough description of the effects of the policy change. I just note that, largely as a result of the change, inflation declined from 10 percent in 1979 to under 5 percent in 1982, most of the gains being achieved in 1982. At the same time, unemployment rose from 5.8 percent in 1979 to just under 10 percent in 1982.

To summarize, although there was a definite policy change, it was not immediately believed by all agents. Direct, informal evidence suggests that it took more than a year to fully change the beliefs of financial

*Massachusetts Institute of Technology, Cambridge, MA 02139. I thank Stanley Fischer for discussions.

[1] Target growth ranges chosen in July 1979 for 1978:4 to 1979:4 were $1\frac{1}{2}$ percent to $4\frac{1}{2}$ percent for $M1$, 5 to 8 percent for $M2$, 6 to 9 percent for $M3$; they have remained approximately the same since.

[2] In the period following October 1979, $M1$ consistently undershot its target range, while $M2$ and $M3$ overshot theirs.

[3] "Direct evidence" is the set of comments of market participants and analysts found in *Business Week*, for the period October 1979–June 1983.

markets, perhaps more than that to change those of labor markets. Thus, in the next section, I address the question of how the economic structure might evolve in response to such a policy change.

II. The Effects of a Policy Change

The monetary mechanism embodied in the econometric models is an intricate one, going from real money balances to short nominal rates, to long rates and asset prices, to the components of aggregate demand, and finally to unemployment and price movements. I focus on two of the links for which expectations are usually believed to play an important role: the first is the relation between short and long nominal rates, or "term structure" relation; the second is the relation between inflation and unemployment, or "Phillips curve" relation.

Let us think of nominal money growth as following a stationary process around a positive mean and of the policy change as a decrease in this mean.[4] How will the two estimated equations perform after the change? More precisely, will the estimated Phillips curve, given unemployment, under- or overpredict inflation? Will the estimated term structure relation, given the short rate, under- or overpredict the long rate?[5]

Consider first the Phillips curve. Most recent macroeconomic models, whether of the imperfect information or contracting varieties, have the same qualitative implication: estimated Phillips curves (i.e., estimated relations between inflation and unemployment) will overpredict the effects of a fully anticipated movement in money growth on unemployment. If, as in the original model by Lucas, there are no predetermined elements affecting the behavior of prices and wages, and if the policy change is instantaneously believed, inflation will decrease with little movement in unemployment. The Phillips curve will therefore, given unemployment, consistently overpredict inflation. If, however, prices and wages depend on past decisions or anticipations, the tradeoff becomes more favorable as time passes, and decisions or anticipations formed before the policy change play less and less of a role. The quality of the Phillips curve forecasts may then deteriorate only slowly. The same is true if agents do not instantaneously believe the change in policy, but slowly revise their beliefs as they observe lower average money growth. In the limiting case where agents do not believe the policy change at all, they do not change the way they form expectations, and the quality of the Phillips curve forecasts may remain the same as before the change.

Turning to the term structure, what happens is more ambiguous. Assume that the movement of short and long rates is determined by the expectations hypothesis, so that the long rate is approximately a weighted average of current and expected future short rates. As in general, expected future rates move less than current rates, long rates move in the same direction as, but by less than, short rates; this is what is captured by empirical term structure equations. Suppose now that there is little predetermination in prices and that the policy change is believed by both labor and financial markets; then, although real money balances may be temporarily lower and the short-term nominal rate higher, anticipations are of lower inflation and of lower nominal rates in the future. The long-term nominal rate might well decline as the short rate increases; the estimated term structure will, given the short rate, overpredict the long rate. Suppose, on the other hand, that prices and wages are largely predetermined, or that labor markets are less convinced of the existence of a policy change than financial markets. Financial markets will anticipate inflation to decrease only slowly and real money balances to be lower for a sustained period of time. They will anticipate nominal rates to remain high for a long period of time: long rates may increase by nearly as much as short rates. The estimated term structure will, given the short rate, underpredict the long rate. To summarize, whether the term structure under- or overpredicts long rates depends on the speed at

[4] This characterization ignores the second aspect of the policy change, that is, the change in the feedback rule.

[5] A formal model is developed in the working paper version of this article.

which financial markets expect inflation to decline after the policy change.

III. The Phillips Curve During 1979–83

I choose to concentrate on the Phillips curve of the DRI model, as specified and estimated in 1978. It is representative of other wage-price Phillips curves. I have also examined the wage-wage Phillips curve specified by George Perry (1978);[6] the results were similar and are not reported here.

The DRI Phillips curve is specified as[7]

$$\dot{\omega} = \alpha_0 + \alpha_1 \dot{p}_{-1} + \alpha_2 \dot{p}^e + \alpha_3 \log u + \varepsilon,$$

where $\dot{\omega}$ denotes wage inflation, with ω being the BLS earnings index: $\dot{p}_{-1}$ denotes lagged inflation, with p being the implicit price deflator; $\dot{p}_e$ denotes "expected inflation," expressed as a geometric distributed lag of past inflation, with decay coefficient .15; and u is the unemployment rate for married males. The unit period is the quarter.

Table 1 gives the results of estimation as years are added to the sample.[8] There is extremely little change in the coefficients until 1982, thus no apparent direct (credibility) effects of the policy change. There is, from 1982 on, some evidence of an increase of α_1 compared to α_2, that is, a decrease in the mean lag effect of price inflation on wage inflation. There is also, in the last regression, some evidence of a larger effect of unemployment on wage inflation: this is more likely due to the very high unemployment rate in 1983 than to direct policy-change effects. The impression of stability is confirmed by the subsample stability tests reported in the last line.[9]

TABLE 1—THE PHILLIPS CURVE

	Final Quarter of Estimation[a]			
	79:3	81:3	82:3	83.2
α_0	5.2	5.2	5.4	5.5
α_1	.22	.25	.33	.36
α_2	.42	.39	.26	.29
α_3	−1.53	−1.48	−1.56	−2.03
D-W	1.96	2.04	1.97	1.86
SE	.99	1.00	1.01	1.01
F^b		1.9	3.2[c]	.5

[a] In each estimate the first quarter is 64:2.

[b] Test statistic associated with the hypothesis of no change in the last year of the sample; distributed $F(4, x)$, $x = 60, 64, 68$, respectively.

[c] Significant at the 5 percent level.

TABLE 2—ONE-QUARTER-AHEAD FORECAST ERRORS

Quarter	$\dot{\omega} - \hat{\dot{\omega}}$	Quarter	$\dot{\omega} - \hat{\dot{\omega}}$
1980:1	.36	1981:4	−2.57
:2	.38	1982:1	−1.65
:3	−.89	:2	−.42
:4	.70	:3	−1.35
1981:1	.57	:4	−1.18
:2	−1.60	1983:1	−2.03
:3	.31	:2	−2.19

Table 2 gives one-period-ahead forecast errors using actual values of the right-hand side variables, and the equation estimated over 1964:2 to 1979:3. The errors are small until the end of 1981: there is again no noticeable effect of the policy change. The errors are, however, consistently negative from 1981:4 on: actual inflation is less than predicted; three of these forecast errors, including two in 1983, are more than twice the standard error of the regression (forecast errors, using the Perry-type wage-wage equations, are also consistently negative from 1981:4 on.) This might indicate a potential, though belated effect of the policy change.

Overall, there is no evidence of a major shift in the Phillips curve. This in no way implies that the above relation is a correctly specified, structural relation, only that the movement of wage inflation, given unemployment, has not been strongly affected by the policy change. This may be due either to unchanged ways of forming expectations, or

[6] The estimated equation is a quarterly version of equation 5.7 in Perry's Table 5. The lagged wage terms are replaced by a geometric distributed lag of past wage inflation, with decay coefficient equal to .25.

[7] See Otto Eckstein (1983, Table 13.2, p. 208) for a more precise description.

[8] The estimates with final quarter 80:3, omitted because of space constraints, are very similar to those with final quarter 81:3.

[9] A more thorough analysis of the stability of the Phillips curve is performed by Steven Englander and Cornelis Los (1983). They also find little evidence of subsample instability.

to expectations playing little role in the determination of wage inflation.

IV. The Term Structure during 1979–83

I concentrate on the quarterly term structure relation of the 1979 version of the MPS model, which was specified and estimated by Franco Modigliani and Robert Shiller (1973).[10] It is specified as follows:

$$R_L = \alpha_0 + \beta_0 R_s + \sum_{i=1}^{19} \beta_i R_s(-i)$$

$$+ \sum_{i=0}^{19} \gamma_i \Pi(-i) + \delta_0 V + \varepsilon;\ \varepsilon = \rho\varepsilon(-1) + u,$$

where the long-term rate R_L is the yield on Aaa bonds; the short-term rate R_s is the three-month rate on prime commercial paper, Π is the rate of *CPI* inflation, and V is an index of variability of the short rate, measured as an eight-quarter moving variance of R_s. The distributed lag structures are third-degree polynomials.

Table 3 gives the results of estimation as years are added to the sample.[11] There is, except in the last year, no clear change in the coefficients; there is, however, a rapid deterioration of fit. The standard error of the residual increases from 25 to 45 basis points. Subsample stability tests, reported in the last line, show each of the years to be significantly different from previous ones.

TABLE 3—THE TERM STRUCTURE

	Final Quarter of Estimation[a]			
	79.2	81:2	82:2	83:2
α_0	1.16	1.02	1.00	1.49
β_0	.19	.19	.18	.19
$\sum_{i=1}^{19} \beta_i$	.61	.65	.66	.45
$\sum_{j=1}^{19} \gamma_i$	.24	.20	.22	.32
δ_0	.02	.09	.07	.14
ρ_0	.61	.57	.49	.67
SE[b]	25.0	27.8	28.5	45.6
F[c]		8.3	10.9	11.1

[a] Beginning quarter is 54:4.

[b] Standard error of the serially correlated residual ε, in basis points.

[c] Test statistic associated with hypothesis of no change in the last year of sample; distributed $F(4, x)$, $x = 94$, 98, 102, respectively; and are significant at the 5 percent level.

TABLE 4—ONE-QUARTER-AHEAD FORECAST ERRORS IN BASIS POINTS

Quarter	$R_L - \hat{R}_L$	Quarter	$R_L - \hat{R}_L$
1979:4	−8	1981:4	168
1980:1	42	1982:1	159
:2	38	:2	64
:3	37	:3	77
:4	66	:4	−90
1981:1	4	1983:1	−84
:2	102	:2	−28
:3	67		

Table 4 gives one-period-ahead forecast errors, using actual values of the right-hand side variables and the equations estimated over 1954:4 to 1979:2. From 1980:1 to 1982:3, forecast errors are large and positive. Although forecast errors from 1982:1 on may be ascribed to unexpectedly large prospects of fiscal deficits, those from 1980:1 to 1981:4 are likely due to the change in monetary policy. Thus expectations appear to have changed and the term structure is very much subject to the Lucas Critique. The fact that forecast errors are positive suggests that, although financial markets slowly believed the policy change, they did not expect inflation to slow down rapidly nor did they expect labor markets to react to the policy change. This is consistent with the evidence on the Phillips curve presented above.

REFERENCES

Eckstein, Otto, *The DRI Model of the U.S. Economy*, New York: McGraw-Hill, 1983.

Englander, A. Steven and Los, Cornelis A., "The

[10] Shiller, John Campbell, and Kermit Schoenholtz (1983) have also reexamined recently the behavior of this term structure equation. Their results are very similar.

[11] Again, because of space constraints, the estimates for the period ending 1980:2 are omitted. These estimates are very similar to those for the period 1981:2.

Stability of the Phillips Curve and Its Implications for the 1980s," Working Paper, Federal Reserve Bank of New York, January 1983.

Lucas, Robert E., Jr., "Econometric Policy Evaluation: A Critique," in Karl Brunner and Alan Meltzer, eds., *The Phillips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl. 1976, 19–46.

Modigliani, Franco and Shiller, Robert, "Inflation, Rational Expectations and the Term Structure of Interest Rates," *Economica*, February 1973, *40*, 12–43.

Perry, George, "Slowing the Wage-Price Spiral: The Macroeconomic View," *Brookings Papers on Economic Activity*, 2:1978, 259–300.

Shiller, Robert, Campbell, John and Schoenholtz, Kermit, "Forward Rates and Future Policy: Interpreting the Term Structure of Interest Rates," *Brookings Papers on Economic Activity*, 1:1983, 173–217.

Foundations of Aggregate Supply Price

By Otto Eckstein*

The experience of the 1970's showed that variations in aggregate supply price were the predominant influences on the price level. This paper provides an empirical measure of this concept and explores its foundations.

Inevitably, the central issue of macroeconomics—how expectations are formed—enters the discussion.

I. A Theoretical Sketch

The following model, though much simplified, is sufficient to show the theoretical underpinnings of aggregate supply price. There is only one factor of production, no productivity trend, and at some points, a simple monetarist version of the rational expectations theory. Let

$$p = \alpha_1 w + \alpha_2 s - \alpha_3 u \tag{1}$$

where p is price, w is the wage rate, s is the index of supply shocks, and u the difference between the actual and the natural unemployment rate. All variables, except the unemployment rate, are percent rates of change.

Aggregate supply price is the sum of the first two terms, that is,

$$p_s = \alpha_1 w + \alpha_2 s. \tag{2}$$

The wage equation (3) relies on price expectations and unemployment, but leaves room for both forward-looking expectations based on an economic model and assumptions about policy, and backward-looking expectations, such as adaptive expectations where the agents learn from experience. Thus,

$$w = \beta_1 p^e + (1-\beta_1)p_{-1} - \beta_2 u, \tag{3}$$

where p^e is the forward-looking expectation and β_1 is the credibility coefficient (or degree of belief) attached to the model and the announced policy rule. Total expectation is the weighted sum of the forward- and backward-looking beliefs.

In the case of the simplest monetarist-rational expectations viewpoint, $\beta_1 = 1$, and $p^e = m^*$, where m^* is the announced monetary target (with the trend in velocity offset by the economy's natural rate of growth).

Aggregate supply price will be

$$p_s = \alpha_1 m^* + \alpha_2 s - \alpha_1 \beta_2 u. \tag{4}$$

On the other hand, if $\beta_1 = 0$, so that price expectations are formed only from actual experience, aggregate supply price becomes

$$p_s = \alpha_1 p_{-1} + \alpha_2 s - \alpha_1 \beta_2 u. \tag{5}$$

Finally, if price expectations are based on a combination of *ex ante* beliefs and the record of actual experience, aggregate supply price becomes

$$p_s = \alpha_1 \beta_1 m^* + \alpha_1 (1-\beta_1) p_{-1} + \alpha_2 s - \alpha_1 \beta_2 u. \tag{6}$$

Actual price is aggregate supply price as modified by the demand element in the product market, so that, in the general case,

$$p = \alpha_1 \beta_1 m^* + \alpha_1 (1-\beta_1) p_{-1} + \alpha_2 s - (\alpha_1 \beta_2 + \alpha_3) u. \tag{7}$$

To measure aggregate supply price requires estimation of the cost-price relationship of equation (1), including its lag structure and explicit treatment of capital as a factor of production, as well as a wage equation which identifies the mix of mechanisms for price expectations.

*Data Resources, Inc., 24 Hartwell Avenue, Lexington, MA 02173. I am grateful to Robert Gross for very capable research assistance on the price equations, and to Sara Johnson for her work on the wage equations.

II. The Price Equation

Tables 1 and 2 summarize the regression results. The basic price equation, using first-degree distributed lags, relates the quarterly inflation rate to a composite index of shock impulses, changes in unit labor costs, changes in capital costs, and to the rate of change in the unemployment rate.

All the cost variables are highly significant. Only the demand variable is weak, but this is not surprising since the principal vehicle of aggregate demand on price is through the various cost variables, particularly unit labor costs.

The time lags on the shock index and unit labor costs are brief, with an average lag of 0.81 quarters for shocks and 1.45 quarters for unit labor costs. Thus, variable costs enter pricing decisions with an average lag of about one quarter, suggesting that the slowness of the inflation response to changes in the level of economic activity does not originate in price behavior but in wages.

Capital costs affect prices with much longer lags. In equation (1), the average lag on capital costs is 10 quarters, and the maximum is 24 quarters. This is not surprising, given the relationship between capital costs and price in the textbook theory of the firm. In the short run, capital costs are fixed; in the long run, they are variable through investment in new capacity justified by prices which include a sufficient rate of return to attract new capital. The lag between capital costs and price should, on average, be related to the turnover rate of fixed capital. The disequilibrium in the rate of return which varies the capital stock presumably originates in the demand term of the equation.

This result shows that production functions of the Cobb-Douglas type are not an adequate foundation for a price equation. A putty-clay-type of formulation is needed to portray the dynamics by which capital cost enters supply price.

The introduction of a monetary variable improves the equation. The growth rate of the narrow money supply over the preceding six quarters is significant with a t-statistic of 4.1, reducing the coefficients and statistical significance of the cost variables only slightly. It reduces the aggregate demand variable below statistical significance, perhaps indicating that it is, in large part, a demand variable, along with its role as an expectational variable on the supply side of price determination.

Robert Gordon (1980) has advanced an equation which represents "inertial" infla-

TABLE 1—PRICE EQUATIONS[a]

	(1)	(2)	(3)
Constant	.690	.246	.252
	(3.65)	(1.19)	(1.21)
Shock Index	.829	.806	.827
	(9.13)	(9.50)	(7.76)
Labor Cost	.298	.255	.262
	(4.32)	(3.90)	(3.77)
Capital Cost	.260	.194	.201
	(4.65)	(3.57)	(3.43)
Price Controls	−.464	−.675	−.690
	(−1.79)	(−2.73)	(−2.73)
Unemployment	−.645	−.353	−.345
Change	(−2.18)	(−1.24)	(−1.20)
Money Growth		.213	.219
		(4.10)	(3.90)
Lagged Price			−.029
			(−.32)
D-W	1.68	1.98	1.94
$\bar{R}^2$	.876	.892	.891

[a] Period of fit: 1955 to 1983:3; t-statistics are shown in parentheses.

TABLE 2—WAGE EQUATIONS[a]

	(4)	(5)	(6)
Constant	.0081	.010	.0008
	(16.35)	(19.99)	(0.62)
Price Record	.750		
	(18.17)		
Price		.580	
Expectations		(15.07)	
Unemployment	−.0015	−.0008	−0.0016
	(−6.95)	(−4.75)	(−5.46)
Guideposts	−.0021	−.0043	
	(−3.45)	(−6.20)	
Minimum Wage	.021	.018	.024
	(3.61)	(2.64)	(2.40)
Money Growth			1.26
			(11.09)
D-W	1.66	1.14	1.05
$\bar{R}^2$	.837	.755	.651

[a] See Table 1.

tion by introducing the dependent variable, price change, in lagged form, but does not include capital costs. This model was tested by adding a lagged price term to equation (2). It was not significant and had the wrong sign. Thus inclusion of capital costs, with long lags to reflect their gradual entry into price, is to be preferred to the inertial inflation model.

The equations are free of serial correlation and show no bias in the recent observations. Whatever role "surprises" may have played in the current disinflation did not make themselves felt in price determination, but rather in the formation of costs.

III. The Wage Equation

Wage equations are well-established, with price expectations and the unemployment rate accounting for most of the observed variation. In the current context, there are five issues:

1) How large are recent overestimates of traditional equations, that is, has anything changed?

2) What is the effect of unemployment on wages when it becomes very high?

3) Can better equations be obtained by using *ex ante* price expectations reported in surveys?

4) Can monetarist rational expectations provide an explanation?

5) Have there been significant structural changes in the labor market to account for recent surprises in wage behavior?

The Phillips curve has traditionally been drawn with a flat tail, assuming that wage rigidity is a characteristic of depression. The wage equations of the last several years had already abandoned that idea, and switched to a Phillips curve that is a straight line with considerable downward slope in the extreme high-unemployment range. Indeed, the Phillips curve may have an S-shape, with very high unemployment breaking the wage structure; but only detailed industry studies can answer that question because there are few observations of double-digit unemployment in the recent historical record.

Equation (4) in Table 2 shows a typical, albeit traditional, wage equation, fitted through the recent data. It relies on the price record, with mean lag of 4.3 quarters, and unemployment. The results show no measurable changes in structure, confirming earlier results (see Steven Englander and Cornelis Los, 1983). The negative errors only begin in the most recent 2 quarters, with an overestimate of 1.8 percent in 1983:2 and 1.1 percent in 1983:3. The average error of the last four quarters is just 0.6 percent, the average error of the last eight quarters is 0.4 percent. These figures give some indication of the new and surprising component of recent wage behavior.

Equation (5) uses the *Michigan Consumer Survey* data as the measure of the public's price expectations. It shows substantially worse results than adaptive expectations. The statistical quality is lower, and recent errors and their bias somewhat larger. These objective measures of price expectations give no hint of a spontaneous improvement in price expectations.[1]

Equation (6) uses a distributed lag on the growth of the narrow money supply as the only price expectations variable. It is substantially inferior to the traditional equation (4), but has explanatory power. A fourth-degree polynomial on the change in M1, with a mean lag of 11.9 quarters, has a t-statistic of 11.1, and leaves the effect of the unemployment rate intact. However, this equation shows larger and more biased errors, with an average error of 0.8 percent for the last four quarters.[2]

The overestimates of the two most recent quarters indicate a surprisingly low rate of wage change. This "surprise" is a residual to be explained by some systematic factor. In the light of the empirical results reported here, it appears that it is not a change in the expectational mechanism or a change in the believed regime governing the economy, but rather a structural change in the labor market. Heighted international competition has had large and visible effects on collective bargain-

[1] When the *ex ante* measure is added to equation (4), it has no explanatory value.

[2] When the money growth factor is added to the traditional equation (4), it does not show any independent explanatory power.

ing in several of the most prominent industries. The deregulation of the transportation industry, including airlines, trucks, and buses, also has had highly visible effects on wages. Thus, the combination of deep recession, intensified international competition, and deregulation has brought about breaks in the wage structures of a sufficient number of industries to affect the national wage totals.

IV. Conclusion

The measurement of aggregate supply price requires identification of the cost-price relationships and equations for endogenous costs. The results show that such relationships can be established by econometric methods, and that aggregate supply price approximates actual price rather closely. Both labor and capital costs, along with measured shocks, play a major role, with variable costs quickly converted into price with an average lag of about one quarter. Capital cost enters price more slowly, as the theory of the firm would suggest. Direct demand effects on price continue to be hard to find in the behavior of aggregate price indexes, but the growth in the narrow money supply in the preceding six quarters does seem to affect prices directly with a coefficient of 0.2. It remains to be seen whether the direct monetary effect is a measure of effective demand or an expectational measure on the supply side of price setting.

In the case of wages, the traditional equation relying on adaptive price expectations serves well, but does show a bias in the last two quarters. Monetarist expectations add nothing to this equation. Changes in the structure of the labor market created by deregulation and international competition probably account for the two recent observations.

APPENDIX

The price variable is the *GNP* deflator for consumer expenditures. This index was selected because the *Consumer Price Index* was redefined in 1983 and the earlier data were marred by the unsatisfactory treatment of the price of homeownership. The *GNP* deflator, on the other hand, is affected too much by the inclusion of components that are difficult to measure, particularly imports and government expenditures. The equation explains the quarterly inflation rates expressed in annualized form.

The index of shock impulses is a composite of the five measured irregular sources of cost changes originally identified in my core inflation model (see my 1981 study). They are the world price of oil, the wholesale price of raw agricultural commodities, the exchange rate, the payroll tax rate, and the minimum wage. The direct shock impulses were originally calculated from simulations of the DRI Model of the U.S. Economy, but were used here without a lag structure in order to let the regression identify the speed of adjustment. Not surprisingly, the time profile is similar to that found in model simulations.

Unit labor cost is measured by the index of total compensation per man-hour in the nonfarm economy divided by a five-quarter moving average of productivity. The averaging of the productivity variable removes the effects of quarterly productivity changes which business seems to ignore in its pricing decisions. On the other hand, it is a shorter moving average than the measures of the studies of the 1960's, when the difference between standard and actual labor cost was sufficiently small and short lived for business to be pricing on cost measures standardized on normal rates of operation.

Capital cost is measured by a Jorgensonian rental price of capital for nonresidential fixed investment, a weighted average of the rental prices of producers' durable equipment and nonresidential construction. The elements of the rental price are long- and short-term interest rates, the cost of equity capital, corporate tax rates, investment tax credits, and depreciation lives. The costs of the sources of capital are weighted by their relative significance in the financing of investment.

A measure of the effects of aggregate demand is necessary to complete the price equation. The conventional demand measure that proved statistically most significant was the change in the deviation between the actual rate and the natural rate of unemployment, where the natural rate was defined in demographic terms.

The growth in the money supply is the annual rate of increase over the preceding six quarters. The only dummy to prove significant represents the Nixon price controls defined to reflect the various "phases" of the program, and constrained to sum to zero by the end of the decontrol period.

The wage equations are fitted in log change form. The independent variables are a third-degree distributed lag on the deflator for consumer expenditures, the difference between the actual and the natural unemployment rate, the change in the minimum wage, a guidepost dummy set at one, 1964–67, the *Michigan Survey* response to the question of expected price change for the next twelve months, and the rate of growth of the narrow money supply.

REFERENCES

Eckstein, Otto, *Core Inflation*, Prentice-Hall, Inc., 1981.

Englander, A. Steven and Los, Cornelis A., "The Stability of the Phillips Curve and Its Implications for the 1980's," Research Paper 8303, Federal Reserve Bank of New York, February 1983.

Gordon, Robert J., "Price Inertia and Policy Ineffectiveness," Working Paper, National Bureau of Economic Research, 1980.

LIMITS TO GROWTH: RESOURCES, PRODUCTIVITY, AND SOCIAL ASPECTS

Sources of Reasonable Hope for the Future

By KENNETH E. BOULDING*

Hope is a complex concept, but in all its various meanings it implies optimism about the future. For a reasonable hope, therefore, we have to examine critically the reasons for pessimism, which in its extreme form is despair.

There are perhaps four major sources of pessimism in the world today, each of a somewhat different quality. The first and overwhelmingly the greatest source is the positive probability of nuclear war. The second is resource depletion—the fact that the accumulated stocks of fossil fuels, ores, and even soils, are being used up and diminishing. The third is somewhat related to the second, which is uncontrollable population increase, which diminishes the stock of available resources per capita. The fourth is a group of possible catastrophes which might be categorized as "system breakdowns."

Of all these, the most dangerous by orders of magnitude is the positive probability of nuclear war. The unilateral national defense organizations of the world, such as the U.S. Department of Defense, its Russian equivalent, and so on, may have the capability of putting an end to the evolutionary process on earth. The systems we are dealing with here are so fantastically unfamiliar that we cannot be sure of anything, but the theses of Jonathan Schell and of Carl Sagan that a nuclear war would create either radiation that would destroy all higher forms of life, or a dust cloud that would cover the earth in a pall of almost total darkness for enough years to destroy nearly all life, is a future that has to be assigned some positive probability, however low. It is a very fundamental proposition that any event with a positive probability will happen if we wait long enough. An event, for instance, like a 100-year flood, which has an annual probability of 1 percent, is virtually certain in 400 years. An event of a probability of only one per thousand in any one year is virtually certain in 4,000 years. What the annual probability of a nuclear war is, of course, nobody knows, but it has certainly been increasing in the last few years, and there is no doubt that it is positive. This means that national defense as a system has simply broken down at a system break, and the future will not be like the past, when national defense, at least if it was well managed, provided a fair probability of national security.

There are really two models here. One is the model of deterrence, which we have followed until recently and which has certainly been stable for nearly 35 years. The very nature of deterrence, however, implies that it cannot be stable in the long run. It must have a positive probability of breaking down. If the chance of nuclear weapons going off were zero, they would not deter anybody and it would be the same as not having them. Historically, deterrence has always broken down into war unless the system has changed into something that is not deterrence.

The other theory is that we can have a limited nuclear war which would result in the domination of the victor, who could then put an end to nuclear weapons by world domination. This seems to be the theory of the present Reagan Administration. Quite apart from the fact that even a limited nuclear war would involve probably the total devastation of Europe or of Korea, or wherever it took place, one looks for the institutions which would limit it and finds they are missing. There is a strong tendency for a unilateral

*Institute of Behavioral Science, University of Colorado, Boulder, CO 80309.

national defense organization in danger of defeat to escalate its destructive power to the full. Even a "tit-for-tat" strategy is likely to go on titting and tatting until all is destroyed. The fact that the communications system is likely to be one of the first casualties of a nuclear war means that there would be nobody to stop it. Decision making would be fragmented and people would tend to do what they are ultimately trained to do—that is, fire the weapons.

We can safely say that if the present system continues we are doomed. One of the most secure propositions about the future is that San Francisco will be destroyed by an earthquake in x years; we just do not know what x is. An exactly parallel proposition is that the United States and the Soviet Union will destroy each other, and quite possibly the whole human race, in x years if the present system continues. We will look later at what might happen if the present system does not continue.

The next source of pessimism is resource exhaustion. This involves two factors: the exhaustion of stocks of things that are valuable to us, like fossil fuels, ores, and soils; and the continued increase in the stocks of things that have negative value—that is, pollution, toxic wastes (both nuclear and nonnuclear), carbon dioxide in the atmosphere, and so on. The whole process of the modern world involves the turning of useful substances into at least potentially harmful substances. This is something that obviously cannot go on forever, or even for very long, by the standards of human history. Cheap oil will certainly be gone in 100 years, coal perhaps in 300 or 500; many ores and deposits of useful chemicals like phosphorus may be gone even before that. On the pollution side, the grosser forms of atmospheric pollution are fairly manageable, and in many places have been rather successfully managed. One looks at the carbon dioxide problem, however, which is virtually insoluble, and which could produce profound changes in the whole climate of the earth within a matter of 100 or 200 years if we continue to burn up the fossil fuels. There are profound uncertainties as to what it might do, but the probability that the results would be in some sense catastrophic in terms, for instance, of a rise in the ocean levels, the desertification of large areas that are now agriculturally productive, and so on, cannot be lightly dismissed.

The third source of pessimism is the uncontrollability of population expansion. The population explosion of people of European descent really began in the mid-eighteenth century with a sharp reduction in infant mortality, probably the result of better nutrition. The current population explosion, largely in the tropical countries, is partly a result of the development of DDT and the control of malaria, which resulted in a very sharp decline in mortality about 1950. There always seems to be a considerable lag between a decline in mortality and a corresponding decline in fertility which will offset it. In Britain, indeed, this gap was more than 100 years. Fertility is now declining all over the world, but at a rather slow rate, so that the population explosion is still proceeding very alarmingly. Countries as diverse as Mexico, Egypt, Bangladesh, and Kenya face an expansion of population in the next few decades which could easily create a very hazardous situation for human welfare. The Malthusian nightmare still haunts us, and is what I have elsewhere called the "utterly dismal theorem"—that if the only thing that can check the growth of population is starvation and misery, any technological improvement or expansion of the human niche will simply mean eventually that we have a larger number of humans living in misery than before.

The fourth concern is some kind of system breakdown in social systems, which can happen on a small or a large scale. We see this at the moment in the appalling tragedy of Lebanon, where the disintegration of the society into warring factions is destroying the very fabric of social life. We see the same thing on a much smaller scale in Northern Ireland, on a larger scale in the tragedy of Cambodia, and so on.

These breakdowns can be economic as well as political. The Great Depression of the 1930's in the capitalist world came very close to a cliff in 1932 and 1933, with unemployment rates running 25 or 30 percent, profits negative, real interest still about 3 percent, an almost total collapse of gross private domestic investment, and both fiscal and monetary policies that, on the whole, intensified the crisis. In the United States in those years interest went to 11 percent of the national income, whereas profits were about minus 3. When an employer hires somebody, the employer sacrifices the interest that could have been gotten on the money spent on the wage in the hope of profit on the product of the work. In 1932 and 1933 it was almost literally true that anybody who gave employment was bound to lose by it, and was either a philanthropist, a fool, or a creature of habit. One suspects, indeed, that only habit kept the economy together and prevented unemployment from going to 50 or even 75 percent of the labor force, and the whole economy collapsing.

In the same period, the Soviet Union saw an even more appalling disaster, the First Collectivization of agriculture under Stalin, in which five or six million people died, there was widespread famine and starvation, especially in the Ukraine, ironically enough the breadbasket of the Soviet Union. This was followed by a terror in which the penalty for speaking truth was frequently death. The Soviet Union has still not recovered from that disaster, especially in agriculture.

In the West, again, a byproduct of the Great Depression was the rise of Hitler and World War II. The Hitler vote in Germany is almost completely correlated with the level of unemployment. He rose to power not because of his lunatic racial policies, but because he promised a way out of the Great Depression.

So much for pessimism. What, then, are the grounds for reasonable hope? Hope, fortunately, does not depend on certainty, or people would certainly not buy lottery tickets. It rests on an optimistic image of the future which is perceived as having positive probability (greater than zero). There is reasonable hope for the abolition of war itself as a monstrous perversity of human activity. This, too, is a future that has a positive probability and the very dramatic nature of the problem increases that probability. Here we can draw a certain amount of hope from the experience of the past. There have been ancient institutions which have disappeared, or virtually so, when the system changed to the point where they became intolerable.

There are three possible examples. One is slavery, which disappeared under the impact of a rising moral feeling that human beings were an illegitimate form of property, and also because of the rising economic advantages of free labor in the labor market. Another example is the disappearance of the feudal baron with his castle and his military retinue after the invention of gunpowder and the efficient cannon, which made the castle indefensible and certainly led to the rise of the national state. A third example is the disappearance of duelling in the first decades of the nineteenth century, largely the result of a change in weaponry, from swords to pistols and then to more accurate pistols.

We are now in a situation where the costs of power are much greater than the benefits. The costs of defeat are far less than the costs of war. The cost of victory is far greater than any benefit that might be obtained from it. In the modern world, especially, war is not about real conflicts; it is about imaginary conflicts. There is no real conflict of interest, for instance, between the United States and the Soviet Union. Every extension of the power of one is a benefit to the other. Important evidence for this whole proposition is found in the abandonment of empire, one of the most striking phenomena of the twentieth century. The evidence is overwhelming that the mass of the people of Britain and France gained very substantially from the abandonment of empire. These countries were then able to devote more resources to getting richer rather than devoting these resources to trying to push other people around.

Perhaps the greatest hope for the abolition of war lies in the fact that over a considerable part of the world there are nations between whom it has been abolished. This is the area of stable peace, stretching roughly from Australia to Japan and across North America to Finland. It is a group of some eighteen countries which have no plans whatever to go to war with each other. This stability might not, of course, last forever.

The fact that it exists shows that it is possible, and because of the fact that the payoffs for this are so large as compared with the alternative of nuclear holocaust—almost infinite—it would be very surprising if this mode of relationship does not expand. In that lies our principal hope.

Turning now to the resources pollution problem—which many years ago I referred to as the "entropy trap"—here again there are grounds for reasonable hope. This is particularly true in regard to energy. Even though it is true that our existing society rests on the exhaustion of fossil fuels, there are both renewable sources of energy in the sun and very large potential stores of it in uranium and thorium, and a very large potential store of it indeed in deuterium in the oceans, which would, of course, require fusion.

The greatest source of hope in solar energy is probably solar photovoltaic cells. In the last ten years or so the cost of electricity from this source has fallen from something like 200 times the usual price to something on the order of 10 or 20 times. If this process can continue—of course there is no guarantee that it will—then solar cells may become a very significant source, especially of domestic electricity, in the next few decades.

Another type of approach to the problem of the utilization of solar energy is through "biomass"—that is, growing things to burn. This is almost certainly the first human use of stored solar energy, for what the human race has always used for fire has been wood. It is still an important source of domestic heat and, up to the latter part of the nineteenth century, even fuelled the American railroads.

Wood, however, is not the only combustible. Many plants can be distilled into gasohol, which will run automobiles; as the residue is largely protein, the loss of food supply (which could be serious) is not as dramatic as it might seem. In our present energy crisis, gasohol is far too expensive and must be highly subsidized, as in Brazil. A two- or three-fold rise in the price of gasoline, however, could easily make it competitive. Whether this would be a serious competitor with food supplies is a little uncertain, but we cannot rule out this possibility. Another possibility which has opened up with genetic engineering is that of new forms of vegetation created in the laboratory that would be more efficient at transforming solar energy than the existing forms.

When we come to materials, the outlook is a good deal more gloomy, especially in the long run, than it is for energy. The earth is not a closed system in regard to energy, but it is virtually closed in regard to materials, in spite of occasional accretion from a meteorite. Here the law of conservation really applies—all we can do with materials on the earth is push them around, move them from one place to another. The geological history of the earth produced concentrations of various materials in terms of ores, phosphate deposits, and so, perhaps through the somewhat "anti-entropic" characteristics of plate tectonics, continental drift, and vulcanism. The human race, however, has been very busy for a long time in diffusing concentrated materials—metals and other elements taken from concentrated patches in mines and then spread over dumps or washed out to sea. This diffusion of the concentrated is what Nicholas Georgescu-Roegen has called "material entropy." It is something we certainly have to worry about.

There are two possible offsets here which can give reasonable hope. One is the constant development of substitutes. A few years ago we were afraid of running out of copper. Now with laser beams and plastics it seems almost unnecessary for the transmission of electricity. Indeed, it has become almost a drug on the market. Some of the most significant elements for possible future technologies are extremely plentiful, like silicon, which we are certainly not going to exhaust, and aluminum, although we presumably will frequently have to go to less concentrated, and, therefore, more expensive, deposits.

Economists constantly urge the great potential of the overall payoff structure, which can be modified for social ends by taxes, subsidies, and sanctions. Taxing resources which are plentiful now but will be scarce later leads to early economizing, conservation, and search for substitutes. Taxing pollutants diminishes them. I have even suggested a solution for the population problem by issuing to each person at adolescence marketable licenses to have, say, 2.2 children. These suggestions may be taken more seriously in the future than they are now.

When we come to the problem of potential social breakdowns, prediction becomes extremely difficult, because these do represent extreme positions of systems which are not subject to any simplistic method of prediction, such as trends or projections, no matter how sophisticated. Certainly nobody predicted the Great Depression or Hitler or Khoumeini. Actually it may be more possible to predict cure than illness. We don't really know when disease will strike us, but once it has, we frequently have some ideas about how to cure it, and here also there may be a reasonable hope.

The greatest source of hope is, of course, Julian Simon's "ultimate resource"—that is, the human mind and its extraordinary capacity for learning and the almost irreversible process of accumulation of human knowledge and know-how, and even, one might add, "know-whether"—that is, the structure of valuations. There are very profound unconscious forces in the dynamics of human society and in nature, some of which are benign—the invisible hands beloved of Adam Smith—and some of which are malign—escalating quarrels and arms races, automobile accidents, and the Great Depression. It is the great virtue of the growth of human knowledge that we are able increasingly to bring these unconscious processes into our consciousness and understanding. It is my belief that the "ultimate resource" is a very long way from being exhausted and that the potential of human beings for learning is still very large indeed, which moves me towards optimism and reasonable hope.

REFERENCES

Boulding, Kenneth E., *The Meaning of the Twentieth Century: The Great Transition*, New York: Harper & Row, 1964.

Georgescu-Roegen, Nicholas, *The Entropy Law and the Economic Process*, Cambridge: Harvard University Press, 1971.

Goeller, H. E. and Weinberg, A. M., "The Age of Substitutability," *Science*, February 20, 1976, *191*:4228, 683–89.

Sagan, Carl, "The Nuclear Winter," *Parade Magazine*, October 30, 1983, 3–7.

Schell, Jonathan, *The Fate of the Earth*, New York: Alfred A. Knopf, 1982.

Simon, Julian, *The Ultimate Resource*, Princeton: Princeton University Press, 1981.

Economic Growth, Resource Availability, and Environmental Quality

By V. KERRY SMITH AND JOHN V. KRUTILLA*

It has now been a decade since *The Limits to Growth* by Donella Meadows et al. helped rekindle economists' interest in the importance of natural resource for economic activities.[1] Today, confidence seems to have replaced concern. At least one prominent economist has explained this change by suggesting that erroneous analysis lay at the heart of any conclusion implying economically important natural resources were growing increasingly scarce.[2] Indeed, current books on the subject such as Julian Simon's *The Ultimate Resource* (1981) offer conclusions that are reminiscent of Harold Barnett and Chandler Morse's classic study *Scarcity and Growth* (1963).

Are these judgements warranted? Has research since 1973 provided answers to the questions derived from theoretical analyses of the relationship between natural resources and the maintenance of material well-being? We think not. While these conclusions ultimately may be judged to be appropriate, such a judgement cannot be made on the basis of the present evidence. In what follows we will develop our reasons for this skepticism, considering first the conventional explanations for how resource stringencies have been avoided and what the past decade's research has added; then, in Section II we discuss what has been missed and why it may be important.

I. Have We Learned About Economic Responses to Finite Resources?

The available explanations of how increasing natural resource scarcity had been avoided can all be summarized using the reasons Barnett and Morse adduced to explain their finding that the relative prices and "real unit costs" of natural resource commodities had not increased over nearly a hundred years.[3] These were:

1) When higher grade sources of a resource are exhausted, lower grade sources are found in greater abundance. Moreover, the qualitative difference in various stocks diminish with the lowering of the grades of resources.

2) As a specific resource becomes more scarce, the rate of increase in its price tends to be offset by substitution of other resources. All but the most insistent demands for the resource are thereby reduced or eliminated.

3) Price increases also stimulate greater exploration to locate new deposits and provide incentives for a greater degree of recycling.

4) Technical change is directed toward reducing the costs of providing natural resource commodities, either through reduction in the extraction costs from existing deposits, or the introduction of techniques that make previously uneconomic deposits a part of the economic reserves.

In our judgement, the most appropriate general summary of research on the econom-

*Centennial Professor of Economics, Vanderbilt University, Nashville, TN 37235, and Senior Fellow, Quality of the Environment Division, Resources for the Future, respectively. This paper draws upon research supported by the National Science Foundation under NSF-CER-S7715083 and DAR-79-21805.

[1] Of course, the Arab-Israeli war and emergence of OPEC as an effective cartel a year or so after the book's publication provided tangible stimuli to this concern.

[2] In a speech given at the Annual North American Meeting of the International Association of Energy Economists, Hendrik Houthakker observed that "In a strict economic logic, the world will *never* run out of any mineral. Resource scares are merely errors of analysis" (1983, p. 3). Of course, under ideal conditions the price of a depleting mineral will restrict consumption before it is exhausted. This in itself does not assure that in the real world resource stringencies will be met with smooth adjustment.

[3] The Barnett-Morse real unit cost measure corresponds to the inverse of a total factor productivity statistic.

ics of natural resources during the past decade is that it has enhanced the analytical precision with which the preceding explanations can be described. We know, for example, that given a finite resource stock, per capita consumption (with constant population) can be maintained indefinitely provided there is sufficient ability to substitute a producible input, such as capital, for the finite resource. By contrast, in the absence of resource augmenting technical change, if the elasticity of substitution between capital and the natural resource is less than unity, then consumption must ultimately decline to zero. Of course, this pessimistic conclusion can be reversed provided there is a small, positive rate of resource augmenting technical change in the economy's production activities.

Complete and ideal markets will lead to an efficient intertemporal allocation of an exhaustible natural resource (see Partha Dasgupta and Geoffrey Heal, 1979, chs. 6–8). Moreover, under specialized assumptions, noncompetitive market structures, including monopoly and some types of oligopolistic markets, can lead to efficient extraction profiles. However, the assumptions required for this outcome within noncompetitive markets are stringent and, therefore, unlikely to be fulfilled. Nonetheless, departures from either competitive markets or noncompetitive market structures that would lead to efficient intertemporal resource-use profiles are likely to favor excess conservation rather than too-rapid depletion. Within the set of models of noncompetitive behavior, those formats that introduce limited forms of competition do not always imply extraction profiles that fall between the extremes of perfect competition and monopoly. The formulation of these limited competition models affects the role played by strategic behavior in their respective solutions. This role determines the relationship between the extraction profiles selected and those from competitive and monopoly markets. Incomplete markets affecting the risks experienced by individual firms also lead to a tendency for conservation in intertemporal extraction choices. In general, then, private markets (with complete information) are not likely to lead to excessively rapid depletion of natural resources.

This conclusion seems to imply that past theoretical research provides little basis for public sector concern or involvement in judging (or influencing) the availability of natural resources. However, such a judgement can be misleading. In concluding their comprehensive treatment of the economic theory of exhaustible resources, Dasgupta and Heal observed that even if one assumes there exist extensive low grade resource deposits, economists should nonetheless pay particular attention to the processes involved in allocating exhaustible resources for at least two reasons: (a) efficient use of these resources requires consideration of the rate and timing of depletion of each quality level of the resource and concern for the transitions between them; and (b) recognition of the potential for substitution for exhaustible resources is not the same as certain knowledge of the availability of such substitutes. In the presence of uncertainty, prudence requires explicit consideration of the consequences of exhaustion as one of a set of possibilities facing the economy.

Of course, as Barnett-Morse recognized two decades ago, theory is not the only source of insight into the importance of exhaustible resources to the economy. It should be possible to gauge, based on private market transactions, whether the stringency of natural resources has been increasingly a negative factor in the performance of the economy.

Along with the theoretical models addressing the role of natural resources in aggregate economic activity and the performance of private firms in extracting them, this insight has fostered inquiry into the definition and performance of indexes of resource scarcity. This work has generally rejected Barnett and Morse's preferred index and favored some measure of the Hotelling rent as an ideal scarcity index. Unfortunately, as we elaborate more fully in what follows, the practical measurement of these rents has to date eluded resource economists.

II. What Have We Missed?

We now comment on limitations in the research developed in four areas: the treatment of natural and environmental resources in aggregate models of economic growth; the specification of the role of natural resources in production activities; the available empirical evidence on the viability of Hotelling-type models for describing firm behavior and market outcomes involving natural resources; and attempts to gauge, using both physical and economic criteria, the availability of natural resources.

Nearly all of the formal economic models addressing the importance of exhaustible resources for the maintenance of economic well-being describe decisions in a stylized framework. They characterize how an ideal centralized planning economy, seeking to maximize discounted social utility, would allocate a single exhaustible natural resource optimally. While such models greatly simplify the description of how economic activities use natural resources, in some cases this has been regarded an advantage. That is, to the extent these models permit the fundamental dimensions of a problem to be isolated and these dimensions are found to be important regardless of the nature of the technical details that are present in the real world, then clearly the models have served their purpose.

Unfortunately, recent evidence (see Heal, 1982, and Morton Kamien-Nancy Schwartz, 1982) is calling this premise into question. These models are quite sensitive to the specification of society's objective function and especially to the prospect of one (or more) state variable(s) influencing society's well-being. For example, if the process of extracting or using the exhaustible resource leads to a stock pollutant (i.e., pollution that accumulates in a fashion similar to a capital stock over time) affecting society's well-being, then: the asymptotic behavior of the model's choice variables will be sensitive to initial conditions; there is the prospect for multiple steady-state solutions; and choice variables may exhibit cyclic behavior. These findings imply that the technical details often assumed "away" as part of the development of a simple description of the aggregate use of natural resources *will* matter to the model's results and that these results may not easily generalize. Thus, if these modifications are judged important and relevant, then simplified aggregate models are unlikely to provide a useful basis for policy with exhaustible resources. Unfortunately, little progress has been made (beyond the identification of the problem) to re-address the issue of judging the importance of exhaustible natural resources to the economy as a whole within a framework that incorporates these technical details.

One way of providing such a response involves detailed analyses of the role of natural resources in individual production activities. Prior to 1973, the empirical evidence on the role of natural resources in production activities was almost nonexistent. Since then, there has been an enormous increase in the number of studies (see Ernst Berndt and Barry Field, 1981, for a review of some of this evidence). The available data greatly constrain what can be done. Natural resources are, for example, routinely treated as an aggregate.[4] The pace of introduction of new technologies is not measured. Rather it is assumed to be capable of being approximated by a time trend. In addition, what we can observe is not necessarily relevant to the levels of input usage where resource exhaustion would imply that substitution must take place (i.e., at high capital resource ratios). In these circumstances heat and materials balances are more likely to be an important determinant of the sensitivity of production activities to these constraints.

Both theoretical (Lawrence Lau, 1982) and experimental evidence (Kopp and Smith, 1980, 1983) question the relevance of the available empirical results. Lau's analysis suggests that the conditions for meaningful economywide (or even sectoral) aggregate measures of resources are so stringent that there is little prospect of their being satisfied in actual production activities. Moreover, the experimental results, though much more limited in scope, indicate that aggregation can lead to substantial distortion in the measured substitutions between inputs. These

[4] Recently, Michael Hazilla and Raymond Kopp (1983) have provided the first detailed treatment of resource substitution identifying the constituents of a natural resource aggregate in their neoclassical cost model for the primary metals sector.

experimental findings are equally pessimistic concerning the performance of general indexes of the pace of technological change. Explicit indicators of the introduction of new technologies seem to be required for an accurate description of the effects of these technologies on input usage. Moreover, the available evidence with real world data (see for example, Michael Denny et al., 1981, and Randy Nelson, forthcoming) indicates that there are substantial differences in the description of the factor biases associated with general (i.e., time trend) versus more explicit indicators of new technologies' adoption patterns.

Despite the longstanding interest in the Hotelling model as a description of the behavior of the extractive firm, empirical tests of this model have been quite limited. Heal and Michael Barrow (1979, 1981) found some evidence of arbitrage behavior with both short-term price movements and long-run data. However, the relationships were not consistent with those implied by theory. Smith's (1981) analysis was even less encouraging for the viability of simple forms of the Hotelling framework. Nonetheless, these initial studies can be questioned because of their incomplete data and the corresponding need to use proxy measures for important variables. Recently, Scott Farrow (1983) has substantially improved on the data available for these earlier efforts by using proprietary information from a single resource extraction firm. After conducting the most thorough examination of the Hotelling model's implications to date, his conclusions are similar to those of earlier authors. Thus, at present, we do not have a good explanation for the Hotelling model's poor performance. Of course, Farrow's one firm over a single time period could well be an outlying case. Nonetheless, the empirical relevance of the framework is worthy of serious reconsideration. Our conclusions concerning the role of public policy in responding to the issues posed by Dasgupta and Heal, rely on firms behaving in accordance with some variant of a Hotelling framework.

Interpreting the economic significance of empirical findings is a process that necessarily involves judgement. Theory may suggest the presence of certain regularities (for example, the movements in resource prices net of extraction costs over time should be associated with the rates of return to assets comparable in risk and liquidity to the resource deposits). While the theory can be used to establish the general nature of this association, it relies on assumptions. Interpretation of empirical findings requires judgements on the correspondence of real world processes with these assumptions and the anticipated "strength" of the empirically measured relationships. Some economists might judge any association (however weak) in this case as support for the theory. Others will not. Most economists would probably adhere to the view that the market discipline is strong, and over time with repeated experience, we can expect to observe the predicted economic responses to resource stringencies. Economic agents will react as theory would suggest, though this may only be observable in qualitative terms and may not be evident from short-term observations of behavior. This view is completely consistent with the strategy adopted by Barnett and Morse. Ideally, a Hotelling rent provides the most appropriate signal of the scarcity of a natural resource. These rents are not readily measured directly, and indirect measures rely on the technical assumptions that are often suspect for some of the reasons we have identified. Consequently most current analyses have relied on the relative prices of natural resources to gauge their scarcity. In the most recent of these efforts, Margaret Slade (1982) finds clear evidence of increasing relative prices for eleven of the twelve minerals she considers over the period 1870 to 1978. However, the pattern is one of initially declining relative prices with an upturn that occurs primarily in the first and second quarters of the twentieth century, depending on the mineral.

Thus, even the empirical record we have available which largely reflects the private costs of obtaining natural resources, does not support dismissing resource scarcity. If we add to this the social costs of pollution (including both "static" externalities and stock pollutants) then confident judgements on the long-term maintenance of economic well-being with a constant or growing population, in the presence of finite natural and environmental resources, seem unwarranted based on what we know today.

REFERENCES

Barnett, Harold T. and Morse, Chandler, *Scarcity and Growth: The Economics of Natural Resource Availability,* Baltimore: Johns Hopkins University, 1963.

Berndt, Ernst R. and Field, Barry C., *Modeling and Measuring Natural Resource Substitution,* Cambridge: MIT Press, 1981.

Dasgupta, Partha S. and Heal, Geoffrey M., *Economic Theory and Exhaustible Resources,* Cambridge: Cambridge University Press, 1979.

Denny, Michael et al., "Estimating the Effects of Diffusion of Technological Innovations in Telecommunications: The Production Structure of Bell Canada," *Canadian Journal of Economics,* February 1981, *14,* 24–43.

Farrow, Scott, "An Empirical Method and Case Study to Test the Economic Efficiency of Extraction from a Stock Resource," Working Paper 83–18, School of Urban and Public Affairs, Carnegie-Mellon University, May 1983.

Hazilla, Michael and Kopp, Raymond J., "A Factor Demand Model for Strategic Nonfuel Minerals in the Primary Metals Sector," Quality of the Environment Division, Resources for the Future, November 1983.

Heal, Geoffrey M., "The Use of Common Property Resources," in V. K. Smith and J. V. Krutilla, eds., *Explorations in Natural Resource Economics,* Baltimore: Johns Hopkins University, 1982, ch. 3.

______ and Barrow, Michael M., "The Relationship Between Interest Rates and Metal Price Movements," *Review of Economic Studies,* January 1979, *47,* 161–82.

______ and ______, "Empirical Investigation of the Long-Term Movement of Resource Prices: A Preliminary Report," *Economics Letters,* No. 1, 1981, *7,* 95–103.

Houthakker, Hendrik, S., "Whatever Happened to the Energy Crisis?," *Energy Journal,* April 1983, *4,* 1–8.

Kamien, Morton I. and Schwartz, Nancy, L., "The Role of Common Property Resources in Optimal Planning Models With Exhaustible Resources," in V. K. Smith and J. V. Krutilla, eds., *Explorations in Natural Resource Economics,* Baltimore: Johns Hopkins University, 1982, ch. 2.

Kopp, Raymond J. and Smith, V. Kerry, "Measuring Factor Substitution with Neoclassical Models: An Experimental Evaluation," *Bell Journal of Economics,* Autumn 1980, *11,* 631–55.

______ and ______, "Neoclassical Modeling of Nonneutral Technological Change: An Experimental Appraisal," *Scandinavian Journal of Economics,* No. 2, 1983, *85,* 127–46.

Lau, Lawrence J., "The Measurement of Raw Materials Inputs," in V. K. Smith and J. V. Krutilla, eds., *Explorations in Natural Resource Economics,* Baltimore: Johns Hopkins University, 1982, ch. 6.

Meadows, Donella H. et al., *The Limits to Growth,* New York: Universe Books, 1972.

Nelson, Randy, "Regulation, Capital Vintage and Technical Change in the Electric Utility Industry," *Review of Economics and Statistics,* forthcoming.

Simon, Julian, *The Ultimate Resource,* Princeton: Princeton University Press, 1981.

Slade, Margaret E., "Trends in Natural Resource Commodity Prices: An Analysis of the Time Domain," *Journal of Environmental Economics and Management,* June 1982, *9,* 122–37.

Smith, V. Kerry, "The Empirical Relevance of Hotelling's Model for Natural Resources," *Resources and Energy,* No. 2, 1981, *3,* 105–17.

Will Productivity Growth Recover? Has It Done So Already?

By MARTIN NEIL BAILY*

The effect of the collapse in productivity growth that has taken place in the U.S. economy since about 1965 has been dramatic. If multifactor productivity growth had continued at the same rate after 1965 as before, then nonfarm business output would have been almost 20 percent higher by 1981 than it actually was, with no additional capital and labor being used. In this paper I will review the latest information on productivity and the alternative explanations of the slowdown, with an eye on whether the explanations predict a resumption of growth. I will then suggest that productivity growth is showing some signs of recovery.

I. Productivity Measures, 1950–81, and the Effect of the Cycle

The Bureau of Labor Statistics has recently released estimates of multifactor productivity for the U.S. economy, as well as its traditional labor productivity measures. The index of multifactor productivity is computed by taking an index of real value-added and dividing it by an index of capital and labor inputs. Labor productivity is computed as real value-added divided by labor input alone. Separate productivity indexes are constructed for the private business sector, the nonfarm business sector and manufacturing. The nonfarm business sector provides the best broad productivity measure, and manufacturing is worth looking at also, because it has relatively high quality data.

Part of the deterioration of productivity growth that shows up in the BLS data after 1965 is a result of changing cyclical conditions. Cyclically adjusted productivity growth rates were computed by running the following regression for each of the four productivity measures over the period 1950–81.

$$(1)\qquad \ln X_t = a + b(t) + c_0U_t + c_1U_{t-1} + c_2U_{t-2} + \varepsilon_t .$$

In this equation, X is the productivity measure, $b(t)$ is a time trend plus time-shift dummies (a linear spline), and U is the unemployment rate adjusted for demographic change.

The equations showed the following pattern. The coefficient c_0 is negative and indicates that a one percentage point increase in the unemployment rate is associated with about a one percentage point decline in the level of multifactor productivity in the nonfarm business sector. This falls to only 0.3 percentage points for labor productivity. The corresponding figures for manufacturing are about two-thirds the absolute values of the coefficients for nonfarm business.

The coefficients on the lagged unemployment rates indicate that if unemployment were to rise and remain high, the long-run effect on productivity would be very small. Persistent slack would give a slight boost to labor productivity and cause a slight decline in multifactor productivity. This result makes sense, because high unemployment usually coincides with some unused capital. This depresses multifactor productivity, but means that production is concentrated on the most efficient machines, and so raises labor productivity.

Based upon the regression results, cyclically adjusted productivity series were calculated, and the adjusted productivity growth rates over various periods are shown in Table 1. The table shows the severity of the slowdown even after cyclical adjustment. Multi-

*The Brookings Institution, 1775 Massachusetts Avenue, NW, Washington, D.C. 20036. A longer version of this paper with figures and additional references is available upon request. Alice Keck and Harry Appelman provided valuable research assistance. The views expressed here are my own and should not be ascribed to the trustees, or other staff members of the Brookings Institution.

TABLE 1—CYCLICALLY ADJUSTED PRODUCTIVITY GROWTH RATES[a]

	1950–65	1965–73	1973–81
Labor Productivity			
NonFarm Business	2.48	2.14	0.55
Manufacturing	2.75	2.77	1.52
Multifactor Productivity			
NonFarm Business	1.77	1.17	0.19
Manufacturing	2.10	1.96	0.76

Source: Computed by the author as described in the text.

[a] Shown in percent.

factor productivity growth fell 0.6 percent after 1965 and another 1.0 percent after 1973 in the nonfarm business sector. The slowdown in manufacturing came later, but was almost as severe. Labor productivity growth in manufacturing was sustained after 1973 by rapid capital formation in that sector.

Many people believe that the cycle somehow explains much more of the slowdown than these regression results indicate. I wonder if this belief is fueled partly by wishful thinking. Advocates of expansionary policies would like to pin as much damage as possible on recessions. But there is little evidence for any magnified effects of the cycle. The U.S. economy managed to return to its earlier productivity trend as the economy recovered from the Great Depression. The period of slack from 1958 to 1962 did not apparently pull productivity down in the longer run.

A simple comparison may be helpful on this issue. Nonfarm business output grew by 22.9 percent over the four-year recovery period 1961–65. Increased productivity accounted for 53.3 percent of this increase in output, with the remaining 46.7 percent coming from increased capital and labor inputs. There was a similar recovery over the four-year period 1975–79, when output grew by 20.9 percent. Thus real demand growth was pretty much the same in both periods. But in the latter period, increased productivity accounted for only 29.7 percent of the output increase, while capital and labor inputs contributed the remaining 70.3 percent. Something very different was happening in the economy during the 1970's.

II. Short- and Long-Term Causes of the Slowdown

The best way to decide whether or not productivity growth will recover is to figure out what caused the slowdown. If the cause of the slowdown is something that has ended, then normal productivity growth should resume. If the cause is something that has reversed, then above normal productivity growth should occur. If the cause results from a long-term or permanent change, productivity growth will remain low. This section will review the leading causes of the slowdown, the evidence for them, and whether they have ended, reversed, or are long term.

A. *A Decline in Innovation*

There are two main pieces of evidence for a decline in innovation. The first is that the rate of issuance of new patents has declined. The U.S. patent office granted a peak of 56,000 patents in 1971 to U.S. nationals. This had fallen to 37,000 by 1980. The second piece of evidence is that there has been some decline in *R&D* spending. The ratio of *R&D* spending to *GNP* has declined from about 2.9 percent in the 1960's to a low of 2.2 percent in 1978.

This evidence is far from definitive. Patents measure inventions, few of which actually become innovations that succeed in the marketplace. The rate of patenting is not a very good measure of innovation. Moreover, there has been a trend away from patenting by business, because patenting is costly and is often of little value.

The decline in *R&D* spending resulted mostly from the decline in defense spending. The ratio of civilian *R&D* spending to *GNP* in the United States did not decline. It was higher in the 1970's than in the 1960's. Military *R&D* has some impact on measured productivity through spillover effects—but not much impact.

If, however, a decline in the rate of innovation is an important cause of the slowdown—if we are running out of ideas—then this is a long-term situation and it means that productivity growth is unlikely to recover.

B. *Energy Prices*

The worldwide increases in energy prices that took place in 1973 and 1979 are obvious candidates for explaining the slowdown. And businesses have reacted to higher energy costs. There has been a sharp decline in the trend of energy use relative to capital and labor use since 1973.

The main argument against the importance of energy to the slowdown is that energy costs are a rather small fraction of total costs. The loss of output associated with the slowdown seems too great relative to the energy saved. There have been various attempts to get around this argument, none of them fully successful. The most forceful attempt comes from models in which the loss of output is an initial investment that then yields a long-term stream of energy savings. If correct, these models predict that once energy prices stabilize, above normal productivity growth should prevail.

C. *Inflation*

High and variable rates of inflation, it is argued, cause bad economic decisions, distort the allocation of resources, and so reduce productivity. The problem with this hypothesis is that it is very hard to model a persistent decline in productivity growth resulting from inflation. Most decisions are reversible, so that the high and variable rates of inflation in the 1970's would not have resulted in relative wages and prices that became cumulatively more and more out of line.

If inflation really has been an important reason for weak productivity, then this says that once inflation is conquered, the economy should go back to its original trend *level* of productivity, not just back to the old growth rate. There would be very rapid growth indeed for quite a while.

D. *Regulation*

If regulation intensity is measured by new laws, or pages in the Federal Register, or expenditure by regulatory agencies, then the 1970's saw an explosive increase in regulation. The impact on productivity is very hard to determine. Most estimates find it to be small. Once the flow of new regulatory measures stops, productivity growth should resume. Since the Reagan Administration has pretty much stopped the flow of new regulations, this hypothesis does suggest a growth recovery.

E. *Work Effort*

This explanation of the slowdown says that welfare state transfer programs have eroded work effort. Employers can no longer enforce work discipline because employees no longer care if they are fired. And since the labor input makes up such a large fraction of total costs, any labor-related explanation has a head start. There is little direct evidence available on trends in work effort, so the only approach is indirect. One indirect measure often cited is the upward trend in transfers as a percent of income. The correlation between this measure and the slowdown is not perfect, however. The income cost of losing a job was at a minimum in the late 1960's when unemployment was very low. This is well before the main slowdown. It is quite likely that work effort has been declining for many years, but the work effort explanation of the fairly abrupt and worldwide slowdown after 1973 is a bit implausible.

Do those who propose this explanation of the slowdown see it as temporary or permanent? Some see a deterioration in work effort as part of a general decline of Western society. For them, presumably, the resulting decline in productivity growth is permanent as we slip into stagnation. However, there are those who believe in the direct and immediate impact of transfer payments. They can point to the Reagan Administration's program which has effectively reduced the income support available to the unemployed. Combined with high unemployment, this should have partially restored the threat of being fired.

F. *The Capital Services Explanation*

In my own earlier work I argued that some of the explanations of the slowdown given earlier, and possibly other structural changes in the economy, may have reduced the effective flow of productive capital services from the existing capital stock. For example, the rise in energy costs may have made some energy-inefficient capital obsolete. Regulation has meant that some capital investment has been for nonmarket activities. The expansion of foreign trade has changed the mix of output that can be sold profitably by U.S. producers. In support of the idea I cited the decline in the market valuation of capital and I found that the slowdown has been much worse in the capital-intensive industries. This latter finding not only supports the capital services hypothesis, but also counts against the work effort hypothesis. Another attempt to support my idea was less successful. Not much evidence has turned up so far of large scale scrapping of capital. It may be that the original vintage capital approach that I used was not a sufficiently complex model to capture what has been happening. For example, the combined effects of regulation and high energy costs have pushed utilities toward using some old capital and leaving some new plants unused. This is certainly a loss of capital services, but would not show up as increased scrapping —on the contrary.

The capital services hypothesis says that a series of shocks to the economy disrupted production. It predicts that productivity growth should resume, once the effect of the shocks has passed.

G. *Short Run or Long Run: A Summing Up*

The alternative hypotheses about the slowdown do not make very tight predictions about the future of productivity growth. If we stick to the plausible alternatives, the real choice is between the hypothesis that there has been a depletion of innovation possibilities—this predicts continued slow growth—and the hypothesis that shocks have disrupted either production or the normal pattern of innovation—this predicts a resumption of growth.

III. Has Productivity Growth Resumed?

It is really much too early to tell whether or not productivity growth has resumed. But there is enough data to make it fun to speculate. Up-to-date information on multifactor productivity is not available, but labor productivity is available on a preliminary basis through the fourth quarter of 1983, and it shows very rapid growth indeed. To try and resolve whether this indicates an improvement in the longer-run trend, I have plotted average labor productivity in U.S. manufacturing and nonfarm business for three separate time periods (1957–63, 1973–80, 1980–83) in a way that makes comparisons possible. The peak level of productivity reached at the beginning of each period is indexed to 100 and the plots then show how productivity evolved in subsequent quarters. The resulting figures (available from the author) show the following. Productivity in the manufacturing sector fell further and for longer after the 1973 peak than in other recessions. There was a strong productivity recovery beginning in mid-1975, but the slow trend of growth became evident 1976–79, even though there were no downturns. This contrasted with the 1960's when rapid productivity growth was sustained well beyond the recovery period. The emerging trend since 1980 is strikingly good. *In the three and one-half years since the beginning of 1980, productivity in manufacturing has grown much more than it did over the six and one-half years after 1973.*

The results for 1983 even suggest that the new trend of productivity growth in the 1980's may be faster than the one for the 1960's. This probably will not be borne out over a longer period. It is consistent with the hypothesis that some of the disruptions that adversely affected performance in manufacturing in the 1970's have been reversed and this has resulted in above normal growth so far in the 1980's.

The equivalent plot for nonfarm business also shows evidence of a resumption of productivity growth, but the turnaround is less striking than it is in manufacturing. This indicates that productivity growth in the nonfarm, nonmanufacturing sector remains weak. Based upon data through 1982, it looks as if problems remain in utilities, mining and construction.

Another striking aspect of recent productivity performance is that productivity fell by only a small amount in 1982 despite the great depth of the recession. Some writers have looked at the growth in productivity since the 1982 trough and found it to be weak or normal. This suggests no recovery of trend growth. But this calculation is misleading, because it neglects the fact that businesses were apparently not hoarding labor in 1982. That is one reason unemployment went so high.

IV. Conclusions

The most likely explanation of the productivity growth slowdown is that it was caused by a variety of factors. There may well have been a decline in innovation and work effort that has been going on for some time. But the bulk of the post-1973 slowdown was probably a result of temporary shocks. Some of the causes of the slowdown—notably the energy price increases—were worldwide and resulted in a common slowdown after 1973. Identical policy responses to the worldwide inflation were also a reason why so many countries experienced slow growth at the same time, as cyclical productivity declines were added to the structural decline.

There are signs from the past two or three years that productivity growth in the 1980's will be much better than it was in the 1970's. This is only speculation at present, but if it is correct, then the temporary shocks view of the slowdown gains credibility relative to explanations that suggest a longer-term reduction in productivity growth.

Tax Policy and the Investment Decision

By CHARLES R. HULTEN*

The poor performance of U.S. productivity in recent years has renewed interest in the relationship between tax incentives, capital formation, and economic growth. Particular attention has been given to the impact of inflation on effective tax rates, and to the possibility that an inflation-induced increase in tax burdens contributed to the productivity slowdown through a reduction in the rate of capital formation. Attention has also focused on the role of relative tax burdens in explaining international differences in savings rates and economic growth, and on the extent to which differences in relative tax treatment of various types of capital results in an inefficient allocation of resources.

These issues are the subject matter of the emerging doctrine of "supply-side" economics, which may be loosely defined as the application of microeconomic principles to macroeconomic problems. While supply-side economics is concerned with the response of all inputs to a change in tax policy, I shall restrict attention to the response of investment spending to changes in business tax incentives. I shall first consider how business tax incentives in the U.S. Internal Revenue Tax Code have changed over time, and then provide an overview of the recently developed cost of capital literature on marginal effective tax rates. I will then turn to the question of how these tax incentives have influenced investment behavior. A concluding section offers comments on the relative importance of supply-side effects on economic growth.

I. Marginal Effective Corporate Tax Rates

Effective tax rates are a convenient device for summarizing the tax burden implied by the many complex provisions of the Tax Code. The traditional method of calculating the effective corporate income tax rate involves dividing actual tax payments by total before tax profits. The resulting ratio measures the average effective rate at which corporate income is taxed, and can be used to estimate differential tax burdens across firms and industries. However, because average effective tax rates are calculated with respect to total tax liabilities, they incorporate tax provisions in effect when past investments were undertaken, and consequently do not indicate the marginal effective tax rate on new investment. Since it is the latter that influences the incentive to invest, the marginal effective tax rate is the appropriate measure for discussions of the supply-side effects of taxation.

A method for estimating marginal effective corporate income tax rates has been developed in recent years.[1] It is based on the Robert Hall-Dale Jorgenson (1967) cost of capital model in which the cost of purchasing an asset (q) is equated, in equilibrium, with the present value of the income generated by the asset, net of economic depreciation and taxes. Allowing for debt finance with a level payment mortgage and assuming that taxpayers have sufficient income to absorb all deductions and credits, this relation-

*Senior Research Associate, The Urban Institute, 2100 M Street, NW, Washington, D.C., 20037. I thank Don Fullerton for helpful comments on an earlier draft.

[1] This method was developed by Alan Auerbach and Dale Jorgenson (1980). Surveys of the literature are available in the papers by Auerbach (1983) and Don Fullerton (1984). The latter gives a detailed discussion of the problems inherent in using the average and marginal effective tax rate methods.

ship takes the following form:[2]

$$(1)\quad (1-\theta)q=\int_0^{\infty} e^{-(\delta+\bar{r})s}(1-u)c\,ds + u(1-Bk)\int_0^{T} e^{-(\bar{r}+\rho)s}D(s)q\,ds + kq-\theta qP\int_0^{M} e^{-rs}\,ds + u\theta qP\int_0^{M}[1-e^{-i(M-s)}]e^{-rs}\,ds,$$

where θ is the percentage of q financed by debt, $\bar{r}$ is the real after-tax discount rate, ρ is the expected rate of inflation, δ is the rate of economic depreciation (assumed constant), u is the statutory marginal rate of taxation, k is the rate of the investment tax credit (*ITC*), B is a basis adjustment for the *ITC*, $D(s)$ is the tax depreciation allowed s years in the future, i is the rate of interest on the loan, M is the maturity of the loan, P is the annual level payment per dollar of loan, and c is the user cost of capital, equal to the value of marginal product of capital. Equation (1) can be solved for the user cost to yield the familiar Hall-Jorgenson formula, adjusted for the deductibility of interest payments:

$$(2)\quad c/q=[1-(1-Bk)uz-k-\lambda\theta](\bar{r}+\delta)/(1-u),$$

where z is the present value of the stream of tax depreciation deductions, $D(s)$, as defined by the second integral on the right-hand side of (1). Because tax depreciation allowances are restricted to the original cost of the asset (i.e., are not indexed for inflation), a nominal rate of return $(\bar{r}+\rho)$ is used in the discounting. An increase in the expected rate of inflation will therefore reduce the value of depreciation allowances, an effect much noted in recent discussions of tax policy.

The term λ captures the interest deductibility provision of the Tax Code. For a level payment mortgage, it takes the form

$$(3)\quad 1-\lambda=(1-u)i(1-e^{-rM})/r(1-e^{-iM}) + uie^{-iM}e^{(i-r)M}-1/(i-r)[1-e^{-iM}].$$

If borrowers and lenders are in the same marginal tax bracket, then it can be shown that $r=(1-u)i$ and that $\lambda=0$, implying that the interest deductibility provision does not enter the calculation of the user cost.[3] This is the widely discussed case of "Modified Fisher's Law," and is the condition for the Modigliani-Miller Theorem to hold. When borrowers are in a higher tax bracket than lenders, then $r>(1-u)i$, and the Treasury loses more revenue through interest deductions than is recouped through taxes on interest income. In this case, the interest deductibility provision lowers the user cost of capital.[4]

Marginal effective tax rates can be derived from (2) in a straightforward way. The various parameters of (2) can be measured or imputed by assumption, and implied value of c calculated according to the formula. Since c is the value of marginal product of capital, the before-tax rate of return, h, is defined implicitly by $c=(h+\delta)q$, and the marginal effective tax rate is defined as $u^*=(h-\bar{r})/h$. This definition of the marginal effective tax rate has the intuitively appealing interpretation as the ratio of the marginal tax liability $h-\bar{r}$ to the marginal before-tax income h.

Marginal effective corporate tax rates for structures and equipment are presented in

[2] The formulation used in this exposition is based on my 1983 paper.

[3] When borrowers and lenders are in the same marginal tax bracket (say, 50 percent), the latter require a 10 percent yield to receive an after-tax return of 5 percent. This premium exactly matches the value of the interest deduction to the borrower.

[4] The actual value of λ is the source of great controversy in the effective tax rate literature (for example, Martin Feldstein and Lawrence Summers, 1979, and Jane Gravelle, 1980). When $r>(1-u)i$ and thus $\lambda>0$ (debt matters), marginal effective corporate tax rates are typically negative, implying that the "tax" is actually a subsidy. In theory, the debt ratio (θ) should rise when $\lambda>0$, reflecting the advantage to high bracket borrowers, and thereby increased i until $\lambda=0$.

TABLE 1—MARGINAL EFFECTIVE CORPORATE TAX RATES ON NEW INVESTMENT: 1952–PRESENT

Subperiod	Total	Equipment	Structure
1952–53	60.2	62.3	56.8
1954–63	52.8	53.7	52.5
1964–65	39.5	34.9	47.0
1966–68	36.4	30.9	46.0
1969–70	53.6	54.0	52.8
1971–72	31.3	21.3	48.0
1973–80	32.8	22.6	49.7
1981 Act	4.7	−14.2	36.3
Present	15.8	3.5	36.3

Table 1 for selected subperiods from 1952 to the present. These estimates, derived from my article with James Robertson (1984), assume $\theta = 0$ or $\lambda = 0$, and that $\bar{r}$ is constant at 4 percent. The overall trend in the Table 1 tax rates is downward, with major reductions occurring after the introduction of accelerated depreciation in 1954, the Guidelines and *ITC* in 1962, ADR in 1971, and the Accelerated Cost Recovery System in 1981. The pattern is, however, far from smooth. Tax rates rose in the late 1950's and 1970's due to an increase in the rate of inflation (the latter being mitigated by further tax reduction). The tax rate rose in the 1960's in response to efforts to cool the Vietnam War boom (through a tax surcharge and the elimination of the *ITC*), and was increased in 1982 in an effort to reduce the large federal budget deficits resulting from the 1981 tax cuts.

Table 1 also reveals that corporate structures and equipment were taxed at rather different effective rates. Although not shown, it is also the case that various categories of structures and equipment experienced significantly different tax burdens. Combined with the much higher effective tax rates on nondepreciable assets, which are taxed at the statutory rate (a maximum of 46 percent at present), the results of Table 1 indicate a highly nonneutral tax treatment of corporate capital. This nonneutrality is exacerbated by the treatment of corporate income under the personal income tax (which depends on the method by which corporate income is distributed), and by the tax-favored treatment of owner-occupied housing and municipal bonds.

The marginal effective tax rate model has been extended to incorporate the joint effects of corporate and personal income taxes. Mervyn King and Fullerton (1984) report a declining total tax rate in the corporate sector between 1960 and 1980, with a further decline thereafter. Their study, which allows for a nonzero effect of debt finance ($\lambda \neq 0$), also presents a comparison of effective total tax rates in four countries, and concludes that effective tax rates were lower in the slower growing countries (the United States and the United Kingdom) than in the more-rapidly growing West Germany.

II. Tax Incentives and Investment Behavior

Having reviewed the trend in effective corporate tax rates, we now consider the responsiveness of investment spending to changes in tax policy. Unfortunately, this is one of the most contentious issues in contemporary economics. One view, associated with Keynesian analysis, emphasizes short-run variations in aggregate demand as the primary determinant of investment spending. The supply-side view, on the other hand, assigns little weight to short-run *policy-induced* shifts in aggregate demand, and emphasizes the role of relative prices in transmitting changes in tax policy.

Space limitations preclude a detailed treatment of this controversy.[5] As an overview, it is useful to note that the Keynesian paradigm is one in which the economy is off its long-run production possibility frontier. The capital stock is underutilized, and the policy-induced changes in demand for capital services can be accommodated with existing stocks without the need for additional investment (at least in the early stages of a recovery). Income effects are thus the primary determinant of the change in investment demand (substitution effects play a negligible role), and the indicated policy prescription is to increase aggregate demand through a broad-based tax cut.

[5] A summary of the basic issues in the controversy may be found in Barry Bosworth (1983) and Summers (1981).

In contrast, the supply-side paradigm focuses on movements along the production possibility frontier. Aggregate (Keynesian) income effects are thus absent, and policy changes operate only through substitution effects. Taxes influence investment through movements along the production possibility frontier due to policy-induced changes in relative prices. The emphasis is also on *permanent* policy changes, as opposed to Keynesian stop-go demand management policies, and on specific (as opposed to broad-based) tax incentives.

The most widely used model of investment behavior—the flexible accelerator model—incorporates both supply-side and output effects, and long- and short-run elements. The optimal demand for aggregate or industrywide capital is determined by analogy with the factor-demand equations of the profit-maximizing firm, and the implied demand for investment is the difference between actual and desired levels of capital. Delivery lags give rise to an investment equation which is a distributed lag of output and user cost (as defined in equation (2)). The time horizon over which price and output effects operate is determined empirically by the estimated lags, and the relative strength of the two effects by the sum of the relevant lagged coefficients.

Actual estimates obtained by this model vary greatly with the econometric specification used with the time period considered. The model performs well with data from the 1950's and 1960's, reflecting the investment boom following the 1962–64 corporate tax cuts. However, for the period 1954–78, Peter Clark (1979) has found that the flexible accelerator model (and the closely related model based on "q" theory) does not perform as well as output-oriented accelerator models, and concludes that output was clearly the main determinant of business fixed investment.

While this conclusion has been disputed (for example, Summers), the general disarray of the empirical investment literature should be noted: Feldstein has recently argued that the investment decision is too complex to be captured in formal models, and that "in practice all econometric specifications are necessarily 'false' models" (1982, p. 829). Robert Chirenko and Robert Eisner have investigated the specification of the investment equation in six major U.S. quarterly models and found a large degree of variability. They conclude that "One can get almost any answer one wants by making sure that the chosen model has specifications appropriate to one's purpose" (1983, p. 139).

These empirical problems should not be taken to imply that tax policy is an unimportant component of the investment decision. Expected future output *is* clearly a crucial aspect of a given firm's decision to increase capacity, but *so is expected profitability*. Tax burdens affect net profitability, and a simple example serves to illustrate the potential magnitude of the tax effect. With an empirically plausible value of one for the elasticity of substitution between capital and other inputs, an increase in the investment tax credit from 10 to 12.5 percent would (under current law) increase the demand for capital by 5 percent, other things equal. The 1982 stock of equipment in U.S. manufacturing is estimated by the Bureau of Labor Statistics (1983) to be $247 billion, and the partial equilibrium impact of the *ITC* thus translates into a $12 billion increase in the demand for new real investment. This is a substantial effect when compared to the $33 billion in average real investment for the 1978–82 period, and is roughly the magnitude of the effect associated with the 1981 Tax Act.[6]

The actual impact of tax policy depends on numerous other factors, such as the size of the future federal budget deficit (a particularly important issue at this point in time). It is also important to distinguish between the impact of tax policy on aggregate investment, which depends on response in the aggregate supply of savings, and the impact on individual types of investment, which can be affected even if the supply of savings is invariant to tax policy. Indeed, the large differential in effective tax rates on various types of capital, noted above, suggest that taxes have the potential for exerting a signifi-

[6] Summers also reports a substantial tax effect within the q theoretic framework.

cant allocational effect on investment (and on deadweight loss) without an increase in total saving.

III. Supply-Side Effects: A Summary

The marginal effective tax rate literature does not support the proposition that inflation in the 1970's was associated with a significant increase in effective tax rates (recall Table 1), nor does it verify the contention that international differences in effective tax rates explain differentials in relative capital formation and economic growth. On the other hand, there is evidence that tax policy per se does exert an important influence on the demand for individual types of capital, and that there are gains to be had from a more neutral pattern of effective tax rates.

The supply-side effects on economic growth are difficult to gauge. First, investment and growth is influenced by all aspects of macroeconomic policy, and tax effects may be swamped by other factors (for example, budget deficits). Second, a once-and-for-all tax cut will influence the capital-output ratio, but will not permanently raise the *rate* of growth of capital or output (the desirability of increasing the capital-output ratio depends on Golden Rule considerations). Third, transitory changes in the growth rate of capital stock have a much smaller impact on output growth since the former must be multiplied by capital's income share (approximately .2 for depreciable assets).[7] It is interesting to note, in this regard, that recent BLS estimates indicate that slower capital formation was *not* a significant factor in the post-1973 productivity slowdown (the decline in total factor productivity explains 80 percent of the slowdown in output per hour in the private business economy). The growth rate of the overall capital stock did fall off after 1973 (by about 10 percent), but the growth rate of equipment actually increased. The strength of equipment investment is consistent with the pattern of marginal effective tax rates reported in Table 1, and suggests that tax policy was effective because it *increased*, rather than decreased, the rate of net equipment investment.

[7] This statement must be qualified by the possibility that capital formation may affect output growth through the embodiment of technical change. In this case, capital formation has a double impact on output growth: a direct effect operating through capital deepening, and an indirect effect operating through an increased rate of total factor-productivity change.

REFERENCES

Auerbach, Alan J., "Taxation, Corporate Financial Policy and the Cost of Capital," *Journal of Economic Literature*, September 1983, *21*, 905–40.

______ **and Jorgenson, Dale W.**, "Inflation-Proof Depreciation of Assets," *Harvard Business Review*, September-October 1980, *58*, 113–18.

Bosworth, Barry P., "Capital Formation, Technology, and Economic Policy," Discussion Paper, The Brookings Institution, August 1983.

Chirenko, Robert S. and Eisner, Robert, "Tax Policy and Investment in Major U.S. Macroeconomic Models," *Journal of Public Economics*, 1983, *20*, 139–66.

Clark, Peter K., "Investment in the 1970s: Theory, Performance, and Prediction," *Brookings Papers on Economic Activity*, 1:1979, 73–113.

Feldstein, Martin F., "Inflation, Tax Rules and Investment: Some Econometric Evidence, *Econometrica*, July 1982, *50*, 825–62.

______ **and Summers, Lawrence H.**, "Inflation and the Taxation of Capital Income in the Corporate Sector," *National Tax Journal*, December 1979, *32*, 445–70.

Fullerton, Don, "Which Effective Tax Rate?," *National Tax Journal*, March 1984.

Gravelle, Jane G., "Inflation and the Taxation of Capital Income in the Corporate Sector: A Comment," *National Tax Journal*, December 1980, *33*, 473–83.

Hall, Robert E. and Jorgenson, Dale W., "Tax Policy and Investment Behavior," *American Economic Review*, June 1967, *57*, 391–414.

Hulten, Charles R., "An Analysis of the 167(k) Accelerated Depreciation Program," Working Paper 3167-02-01, The Urban Institute, May 1983.

______ **and Robertson, James W.,** "Corporate Tax Policy and Economic Growth: An Analysis of the 1981 and 1982 Tax Acts," in A. Dogramaci, ed., *Studies in Productivity Analysis*, Vol. 6, Boston: Kluwer-Nijoff, 1984 forthcoming.

King, Mervyn A. and Fullerton, Don, *The Taxation of Income From Capital: A Comparative Study of the U.S., U.K., Sweden, and West Germany*, Chicago: University of Chicago Press, 1984 forthcoming.

Summers, Lawrence H., "The Effect of Economic Policy on Investment," in L. H. Meyer, ed., *The Supply Side Effects of Economic Policy*, Boston: Kluwer-Nijoff, 1981, 115–48.

U.S. Department of Labor, BLS, *Trends in Multifactor Productivity, 1948–81* Bulletin 2178, Washington: USGPO, September 1983.

Family Labor Supply with Taxes

By JERRY HAUSMAN AND PAUL RUUD*

Taxes on labor supply raise the largest proportion of federal tax revenue. Over the period 1960–83, this proportion has increased from 57 to 77 percent. While the income tax has increased moderately as a proportion of federal tax revenue, the payroll tax has more than doubled as a proportion of all taxes. In 1980, approximately 50 percent of federal tax revenues was raised by the individual income tax. The individual income tax is a progressive tax on labor and nonlabor income which is based on the notion of "ability to pay." We finance Social Security by the payroll tax, FICA, which is a proportional tax with an upper limit. As both the tax rate and upper limit have grown rapidly in recent years, FICA taxes have become the subject of much controversy. In 1980, FICA taxes represented 28 percent of total federal tax revenue. It is interesting to note that, between 1960 and 1980, while the marginal income tax of the median taxpayer remained constant, the FICA tax rate more than doubled, and the earnings limit rose about 220 percent in constant dollars. Over the same twenty-year period, the corporate income tax has decreased from 24 to 13 percent of federal tax revenues. Likewise, excise taxes have decreased from 13 to 5 percent. Thus, taxes on labor supply currently amount to about three-fourths of federal taxes raised.[1] The potential effects on labor supply and economic welfare are important because of the large and increasing reliance on direct taxation.

In Table 1, a summary of marginal tax rates for the period 1950–84 is provided. These rates are for married households filing jointly. The *CPI* and median family income are also shown so that valid comparisons across different years can be made. First, note that the tax system between 1950 and 1980 was only imperfectly indexed for inflation. The median income family faced a marginal tax rate of 17 percent in 1950, but multiplied by the change in the *CPI*, this family faced a marginal rate of 21 percent in 1980. Similarly $10,000 of earned income in 1950 had a marginal tax rate of 24 percent in 1950, but adjusted for inflation, the marginal tax rate increased to 37 percent in 1980. Similar increases in marginal tax rates occurred over the periods 1960–80 and 1970–80. Of course, this imperfect indexation corresponds to greater progressivity which may have been the intent of Congress over the period, although the marked increase in tax preference items would lead to lower actual progressivity. Another interesting finding which emerges from Table 1 is the significantly higher marginal tax rates faced by the median family over the period. The marginal tax rate increased from 17 percent in 1950 to 28 percent in 1980. Note that under the tax reform of 1981, marginal rates will drop substantially by 1984 due to the 25 percent tax reduction. Much of the "bracket creep" of the past decade will be eliminated. Under current legislation, the income tax system is scheduled to be indexed beginning in 1985. Also, beginning this year, a 10 percent deduction for two-earner couples up to a limit of $30,000 exists. This deduction reduces the effects of the "marriage tax" which arises due to the progressive structure of the income tax. We estimate the effect of this new deduction on the labor supply of families given our results in the current paper. Appropriate economic tech-

*Massachusetts Institute of Technology, Cambridge, MA 02139 and National Bureau of Economic Research, and University of California, Berkeley, CA 94720, respectively. This research has been supported by the NSF and the NBER. Angus Deaton, Peter Diamond, Mark Killingsworth, and Mervyn King provided helpful advice. David Braver, Allen Cheadle, and Kathryn Rockett provided research assistance.

[1] Of course, not all income tax revenue is a tax on labor supply because of the taxation of capital income which was about 12 percent of adjusted gross income in 1980. Also, a portion of the incidence of FICA taxes fall on the employer, although the amount is likely to be small.

TABLE 1—FEDERAL INCOME TAX: SELECTED MARGINAL RATES FOR MARRIED COUPLES

Taxable Income	1950	1960	1970[a]	1980	1984[b]
2–4	17	20	17	14	11
6–8	20	22	19	16	12
10–12	24	26	23	21	16
16–18	31	34	29	24	18
20–22	35	38	33	28	22
28–32	43	47	40	37[c]	28
40–44	51	56	49	43	33
56–60	56	62	54	49	38
64–70	59	65	56	50[c]	42
76–80	63	69	59	50[c]	42
90–100	66	72	62	50	45
120–140	71	78	66	50	49
180–200	79	87	71	50	50
400^+	84	91	72	50	50
1)[d]	1.0	1.2	1.6	3.4	
2)[e]	3.3	5.6	9.8	21.0	

Note: Taxable income is shown in $1,000s.

[a] Includes 2.5 percent surtax.

[b] The 1984 rates reflect the entire 25 percent tax reduction passed by Congress in 1981. The tax will then be indexed.

[c] Maximum tax on earned (labor) income was 50 percent beginning in 1972 under the Tax Reform Act of 1969.

[d] *CPI* in 1950 dollars.

[e] Median family income in thousands of current dollars.

niques to measure the effect of taxation need to treat the nonlinearity of the budget sets which arises due to the nonconstancy of the marginal after-tax wage.[2]

Recent econometric research has developed techniques which takes account of this nonlinearity and the related situation that the marginal wage rate is jointly determined with the amount of labor supply. In this paper, we extend the recent work to account for the interdependent nature of family labor supply decisions, rather than treating husbands and wives separately. The effect of taxation has potentially important effects here since the labor supply of the husband affects the marginal tax rate of the wife, and vice versa. Fixed costs of work may also have an important role here, especially as they interact with the tax system. Therefore, we combine joint family decisions with the effects of the tax system. When we combine these approaches, an additional theoretical and econometric problem arises. For nonworking spouses, the appropriate "virtual" wage must be estimated since this wage enters the labor supply function of the working spouse. The notion of the virtual wage arises in the theory of rationing and is analogous to the notion of virtual income, previously used in models of labor supply with taxes.[3] That is, the virtual wage is that wage which would cause the individual to choose to work exactly zero hours with a tangency of the family indifference curve and the budget set which is determined by the net after-tax wage of the working spouse, the virtual wage of the nonworking spouse, and the virtual income of the family. The treatment of joint family labor supply with taxation and with the use of the correct virtual wage for nonworking spouses is the main contribution of this paper.

I. Econometric Specification

We have specified a new indirect utility function for the multiple good case. Specifications that have been previously used impose constraints on the supply functions across goods that are improbable in the joint labor supply of husbands and wives. In our sample, almost all of the men work, but only half of the women do. Wives who do work, work approximately one-half as many hours as working husbands. Clearly, the supply behavior is very different among husbands and wives, and the parametric specification must be able to accommodate this.

Our indirect utility function has the following functional form:

$$(1)\quad V(w_1,w_2,y) = \exp(\beta_1 w_1+\beta_2 w_2)\big(y+\theta+\delta_1 w_1+\delta_2 w_2 + .5(\gamma_1 w_1^2+\gamma_2 w_2^2+\alpha w_1 w_2)\big),$$

[2] For recent reviews of research in this field, see Hausman (1984) and Mark Killingsworth (1983).

[3] The use of virtual prices in rationing was introduced by E. Rothbarth (1941). A recent treatment was given by Angus Deaton and John Muellbauer (1980). Note that the virtual wage is not the same as the reservation wage in a model with fixed costs and taxes, compare Hausman (1980).

or, more simply,

$$(2)\quad V(w_1, w_2, y) = \exp(\beta_1 w_1 + \beta_2 w_2) y^*(w_1, w_2, y),$$

where θ, β_1, β_2, δ_1, δ_2, γ_1, γ_2, and α are parameters of the indirect utility function, the w's are the respective net wages, and y is the virtual income. Direct application of Roy's Identity yields the labor supply equations

$$(3)\quad \begin{aligned} h_1 &= \delta_1 + \beta_1 y^* + \gamma_1 w_1 + \alpha w_2, \\ h_2 &= \delta_2 + \beta_2 y^* + \gamma_2 w_2 + \alpha w_1. \end{aligned}$$

These supply equations have the simple form of being linear in virtual income and quadratic in wages, where the quadratic terms appear in y^*. Furthermore, this functional form allows each equation to have its own intercept and income coefficient. The equations are second-order flexible and have the convenient property that nonlinearity arises only in products of coefficients, which makes econometric estimation considerably easier than with other flexible functional forms.

Nevertheless, their derivation from an indirect utility function places constraints on these supply equations. That is, our model of family labor supply assumes maximization of a joint family utility function so that the restrictions imposed by economic theory apply. Because the indirect utility function must be nondecreasing in wages and both compensated supply equations must be upward sloping, despite the quadratic wage terms that allow backward-bending labor supply. Note that although the supply equations appear conventional, the individual parameters in these equations do not have conventional functions of only eight parameters. For example, the derivative of h_1 with respect to w_1 is

$$(4)\quad \beta_1\delta_1 + \gamma_1 + \beta_1(\gamma_1 w_1 + \alpha w_2),$$

which depends on parameters that one might associate solely with income or w_2. Combining the supply equations with the budget constraint

$$(5)\quad x = y + w_1 h_1 + w_2 h_2$$

completely characterizes the behavior of the household. The direct utility function can be derived by solving these three equations for h_1, h_2, and x as functions of w_1, w_2, and y, and direct substitution into the indirect utility function. The direct utility function is needed to predict behavior when one or both spouses are not working, or in the presence of fixed costs. It is convenient to express the direct utility function implicity as follows: given h_1, h_2, and x, both y and w_2 can be expressed as linear functions of w_1. Substituting these linear relationships into the first supply equation yields a quadratic function in w_1 that is easily solved: one chooses the root that corresponds to an upward-sloping supply. Recursive solution for w_2, y, and, finally, utility yields the direct utility function.

The labor supply functions of equation (3) yield the hours that maximize the direct utility function if the budget frontier is linear. This occurs when the net wages are constant for all hours worked and there are no fixed costs incurred by working. But the introduction of taxes leads to a nonlinear budget set.[4] It is necessary to know the direct utility function in order to determine the hours of the husband and the wife predicted by utility maximization over a nonlinear budget frontier, because simple revealed preference arguments cannot determine global maxima in utility. The exception to this rule occurs when the budget frontier is globally convex. The argument in Hausman (1979) continues to work for the multiple good case.

For the tax schedules faced by couples, the maximization of utility can be broken up into maximization over convex subsets of the budget frontier followed by maximization of utility over the entire set of solutions. Each income bracket is a convex subset of the budget frontier that has a relatively simple

[4] In addition, nonconvexities are created by, for example, the standard deduction and income limit on FICA taxes.

utility maximum. Within an income bracket, net wages and, consequently, virtual (nonlabor) income are constant so that locally the budget frontier is the conventional linear budget constraint. This budget constraint must include, however, nonnegativity constraints on hours worked by each member of a couple and income constraints that define the tax bracket.

The utility maximum on such convex sets as an income bracket of the tax schedule does not require the utility function. First, we compute the unconstrained solution using the supply equations. If this solution falls within the income bracket, then this solution is obviously the constrained maximum, too. Otherwise, the utility maximum occurs on a boundary. One, two, or three of the constraints may be violated. Let us examine each case in turn. If only one constraint is violated by the supply equations, then convexity of the direct utility function and the budget set ensures that the constrained solution lies on the corresponding linear boundary. The boundary solution can be found by substituting the boundary constraint directly into the system of supply equations and the budget constraint, and solving for the virtual wages and income in the same manner as we described for deriving the utility function. The constrained solution corresponds to the unconstrained hours supplied at these virtual wages and income.

On a boundary corresponding to zero hours, the wage for the working spouse and the virtual income correspond to the observed ones. Therefore, one solves the quadratic in the virtual wage of the person not working formed by their supply equation to obtain the constrained maximizing point. On the income bracket boundary, the ratio of the virtual wages must be equal to the slope of the hours tradeoff and total consumption is fixed by the total income constraint. These two constraints also lead to a quadratic equation in a virtual wage that leads to the constrained optimal hours. These closed-form solutions are a special feature of our parameterization. Other choices of functional forms for labor supply can lead to extremely difficult systems of equations for which no closed form exist.

If the boundary solutions do not satisfy one of the other constraints, then the constrained solution lies at the vertex where the constraints are simultaneously satisfied. When two constraints are violated by the unconstrained solution, then the constrained solution may rest on either boundary or the common vertex. Again, convexity ensures that only one boundary solution will be feasible. Finally, three constraints will be violated by the unconstrained solution if both supply equations yield negative hours. The constrained solution will lie on the lower income boundary, or one of the couple will not work. One must examine all three cases to find the single, feasible solution.

The stochastic specification which we use allows for a truncated bivariate normal distribution to represent the deviation between actual and desired hours and measurement errors. The truncation arises because of the lower limit of zero hours of work. Therefore, the likelihood function is a bivariate Tobit model where none, one, or both of the stochastic terms may be truncated. An individual may not work either because the preferred hours are zero, or because the realization of the stochastic term is sufficiently negative to induce zero hours. We include fixed costs to working for the wife as in Hausman (1981a), but we do not allow for preference variation through a distribution of parameters.[5] The parameter estimates are unconstrained as we do not impose the global integrability conditions of economic theory beyond those incorporated into the indirect utility function of equation (1).

II. Estimation Results

Our sample is drawn from the 1976 wave of the *Michigan Panel Study of Income Dynamics*. It consists of 1,991 couples that remained after removing the self-employed and farming families, and those observations that were missing data on socioeconomic

[5] The presence of fixed costs to working for the wife are the one asymmetrical part of our specification. However, in the absence of preference variation, they seem to have an important role in the explanation of the difference in labor force participation between spouses.

explanatory variables. We eliminated any household which contained an individual who claimed to work in excess of 4,000 hours per year. We also truncated the sample based on the observed wages, requiring wages of both husband and wife to be less than $20. The missing wages for women who were not working were estimated by a standard sample selection model using the wage data of the working women.[6]

The estimation results for the labor supply functions of equation (3) and the associated indirect utility function of equation (1) are given in Table 2. The method of estimation is maximum likelihood. Person 1 is assumed to be the husband. Note that the coefficients for the virtual income, the own-wage effect, and cross-wage effect are quite precisely estimated. The coefficient for virtual income for husbands equals $-.147$ and is quite close to the mean of the β distribution estimated by Hausman (1981a) where he estimated husbands labor supply independent of the labor supply behavior of wives.[7] The mean income elasticity in the sample for husbands is estimated to be $-.101$ which again demonstrates the importance of taking account of taxes in models of labor supply. Since the net wage enters y^* in equation (3), we use equation (4) to calculate the derivative of hours of labor supply with respect to the wage. The mean derivative equals $-.016$, which corresponds to an elasticity of $-.034$ which is slightly greater than Hausman (1981a) found. The restriction from economic theory that the compensated demand curve be upward sloping is satisfied for hours greater than 231. This restriction is satisfied for almost all the predicted hours which have a mean of 2,140, and for actual hours which have a mean of 2,129. The mean derivative of the cross-wage effect for men is $-.382$. However, the estimate is not significantly different from zero. Nevertheless, it is possible that this finding may arise from the correct treatment of virtual wages for nonworking wives.

TABLE 2—ESTIMATES OF JOINT FAMILY LABOR SUPPLY MODEL[a]

Coefficient	Husband		Wife
δ_j = Constant	1.655		1.090
	(.006)		(.133)
β_j = Virtual Income	$-.147$		$-.316$
	(.008)		(.054)
γ_j = Net Wage	.259		.321
	(.004)		(.132)
α = Spouse Wage		$-.548$	
		(.115)	
θ = Combined Constant		-11.228	
		(.367)	
Ill-Health	$-.084$		$-.507$
	(.059)		(.024)
Children under 6 Years		.001	
		(.004)	
Family Size		.001	
		(.019)	
Fixed Costs			.755
			(.203)
Children under 6 in *FC*			.002
			(.013)
Family size in *FC*			.034
			(.043)

Note: $LF = -9940.$ $\Sigma = \begin{bmatrix} .224 & - \\ .040 & .342 \end{bmatrix}$

[a]Asymptotic standard errors are shown in parentheses; all income variables are shown in $1,000s; fixed costs are shown in $1,000 per year. Hours are used in 1,000's.

The coefficient for virtual income for wives is estimated to be $-.316$ which is about $2\frac{1}{2}$ times larger than the results from Hausman (1981a) where husband's labor supply is taken as exogenous in the determination of the wife's labor supply. The mean elasticity estimate of $-.360$ is quite similar to Hausman's result because we use family virtual income in the elasticity calculation. The average own-wage effect is calculated to be .385 which leads to an elasticity of .757 which is below the previous estimate of .906. However, a substantial own-wage elasticity still exists for wives. The mean cross-wage effect is estimated to be $-.104$ with an elasticity of -2.36. As with the husband's cross-wage effect, this estimate demonstrates the importance of the spouse's wage. It is interesting to note that a 1 percent change in the mean husband's wage leads to a predicted decrease of 30.9 hours for the average wife, while if the wife's response is conditioned on the husband's behavior as in previous Hausman (1981a) models, the predicted reduction

[6]Further details of the sample selection procedure can be found in Hausman (1981a), since similar procedures were employed.

[7]Estimates from other studies are given in Killingsworth.

would be 47.2 hours. The predictions are of similar size, but the response in the joint labor supply model is somewhat less as would be expected.

Since symmetry has been imposed by the specification of the indirect utility function of equation (1), the only other restriction of economic theory is the positive definiteness of the Slutsky matrix. That is, both compensated labor supply curves must be upward sloping, which is a restriction found to be violated in many labor supply studies for men which ignore taxes. Also the compensated cross-wage elasticity cannot be too large. At the mean of the data, both compensated demand curves and the determinant of the Slutsky matrix are all significantly greater than zero when tested at the .05 level. The positivity restrictions are also satisfied at almost all the data points. We conclude that our estimation results do not reject the economic theory of joint household maximization.

We now consider the economic welfare and labor supply effects of the tax treatment of families. Using the indirect utility function of equation (1), we calculate the deadweight loss (DWL) of the taxation of labor income with the approach of Hausman (1981a, b). The ratio of DWL to tax revenues is estimated to be 29.6 percent which should be compared to 28.7 percent for males and approximately 58 percent for wives estimated by Hausman (1981a). Again, we find a rather high cost in efficiency for the progressivity of the tax system. We next consider the effect of the 10 percent deduction introduced in 1983 to reduce the marriage tax.[8] It should be noted that we apply the 10 percent reduction to the 1976 tax rates of our sample rather than the actual 1983 situation. We estimate that wives' labor supply will increase by 3.8 percent, while husband's hours decrease by .9 percent. Overall taxes paid decrease by 3.4 percent.

To compare the desirability of the change on efficiency grounds, we would need to know the DWL of marginal tax revenues raised by other means to compensate for the tax reduction here. But, the equity grounds for a reduction in the marriage tax are quite strong.[9]

III. Interpreting Labor Supply Estimates: The Role of Income Effects

We conclude with comments on two related arguments that sometimes arise in the estimation of labor supply functions. The first argument is that what we measure are actually compensated (Hicksian) labor supply curves, which therefore contain only substitution effects, but no income effects. The second argument is that a change in income taxation will induce only substitution effects so that the change in labor supply has a determinate direction, rather than being composed of potentially offsetting income and substitution effects. These arguments are usually associated with the analysis of Milton Friedman (1949; 1976, ch. 3).[10] The basic idea is that the goods and services which the government provides with the tax revenues will need to be replaced by the private economy when the taxes are reduced. The production possibility frontier of the economy thus remains unchanged. The argument is then made that income effects do not exist in this "general equilibrium" analysis, so that only substitution effects are measured or are of importance in the analysis of changes in tax policy.

Strictly construed, this argument will only hold in a one-consumer economy. In a times-series regression of a so-called representative consumer model, it might be argued that the government expenditure exactly duplicates the individuals' preferences. But this position is untenable in a cross-section regression of nonidentical individuals. In this situation, note that all individuals receive the public goods provided by the government irrespective of their actual labor supply. Therefore, the statistical experiment being

[8] Previous estimates are provided by Daniel Feenberg and Harvey Rosen (1983).

[9] Joseph Pechman (1983) discusses the choice of the appropriate basis for income taxation.

[10] See also Martin Bailey (1954). Recent treatments include Ronald Ehrenberg and Richard Smith (1982), and James Gwartney and Richard Stroup (1983).

undertaken is to compare sample points which have different after-tax net wages and different virtual incomes, but receive identical amounts of public goods. The presence of an income effect is then immediate because of the different levels of utility that the individuals or families reach. The Friedman argument would seem to require that all families are on the same indifference curve which certainly does not hold here. Therefore, the argument that the compensated labor supply curve is being estimated does not stand examination in the typical context of estimation of labor supply with taxation.

The second argument is that only substitution effects are relevant in the consideration of changes in income taxation. The key consideration here involves what is being held constant. Richard Musgrave in his classic treatise discusses the various possibilities and recommends the comparison of "alternative methods of tax financing a given level of real expenditures" (1959, p. 212). Income effects are clearly of importance in this situation. Hausman (1981a) compares the labor supply effects and welfare effects of a switch from the current tax system to an equal yield progressive linear income tax. The level of public goods provision is being held constant, and income effects occur across different individuals as they shift from their current indifference curves to their new indifference curves under the new tax system. The importance of heterogeneity among individuals is ignored in the Friedman-type arguments. The one case where the analysis becomes more difficult is when a tax change with respect to the income tax is being considered without a compensating change in other taxes to hold total revenue constant. But even here, the argument that only substitution effects exist is not correct in general. In fact, in the most recent observation of a major tax change, the level of government expenditure has changed by only a small amount with respect to the decrease in government revenues. Again, income effects would be important in the analysis of the income tax cuts, although a confirmed believer of the Ricardo-Barro analysis might argue that no effective decrease in taxes has taken place because of the concurrent increase in the deficit.

REFERENCES

Bailey, Martin J., "The Marshallian Demand Curve," *Journal of Political Economy*, June 1954, *62*, 255–61.

Deaton, Angus and Muellbauer, John, *Economics and Consumer Behavior*, Cambridge, Coolidge University Press, 1980.

Ehrenberg, Ronald and Smith, Richard, *Modern Labor Economics*, Glenview: Scott Foresman, 1982.

Feenberg, Daniel and Rosen, Harvey, "Alternative Tax Treatment of the Family," in M. Feldstein, ed., *Tax Simulation Models*, Chicago: University of Chicago Press, 1983.

Friedman, Milton, "The Marshallian Demand Curve," *Journal of Political Economy*, December 1949, *57*, 463–95.

______, *Price Theory*, Chicago: Aldine, 1976.

Gwartney, James and Stroup, Richard, "Labor Supply and Tax Rates: A Correction of the Record," *American Economic Review*, June 1983, *73*, 446–51.

Hausman, Jerry, "The Econometrics of Labor Supply on Convex Budget Sets," *Economics Letters*, No. 2, 1979, *3*, 171–74.

______, "The Effect of Wages, Taxes, and Fixed Costs on Women's Labor Force Participation," *Journal of Public Economics*, October 1980, *14*, 161–94.

______, (1981a) "Labor Supply," in Henry Aaron and Joseph Pechman, eds., *How Taxes Affect Economic Behavior*, Washington: The Brookings Institution, 1981.

______, (1981b) "Exact Consumers Surplus and Deadweight Loss," *American Economic Review*, September 1981, *71*, 662–76.

______, "Taxes and Labor Supply" in Alan Auerbach and Martin Feldstein, eds., *Handbook of Public Economies*, forthcoming 1984.

Killingsworth, Mark, *Labor Supply*, Cambridge: Cambridge University Press, 1983.

Musgrave, Richard, *The Theory of Public Finance*, New York: McGraw Hill, 1959.

Pechman, Joseph, *Federal Tax Policy*, 4th ed., Washington: The Brookings Institution, 1983.

The After-Tax Rate of Return Affects Private Savings

By LAWRENCE H. SUMMERS*

The effects of the rate of return on the level of savings and the rate of capital formation are of central concern to both economists and policymakers. Although the welfare effects of tax reforms do not directly depend on their impact on savings, the effects of taxes on savings is crucial to considerations of tax incidence and equity, and to the issue of long-run growth. The impact of the rate of return on consumption and savings decisions also bears on questions regarding the appropriate government discount rate, the short-run crowding-out effects of fiscal policy, and the effects of public indebtedness on capital intensity.

The traditional view among economists is that changes in the rate of return are likely to have only a small effect on the savings rate. This consensus is supported by theoretical arguments pointing to the opposing income and substitution effects associated with changes in the rate of return. The ambiguous implications of theory are matched by empirical studies which yield conflicting estimates as to the size of the impact of changes in the rate of return. The polar empirical estimate is Michael Boskin's (1978) suggestion that the interest elasticity of savings is .4. This estimate is widely regarded as too high.

This paper reexamines the theoretical arguments and reviews new empirical evidence regarding the interest elasticity of savings. Both the theoretical analysis and the empirical work demonstrate the strong likelihood that increases in the real after-tax rate of return received by savers would lead to substantial increases in long-run capital accumulation. While it is not possible to quantify the impact with any precision, it seems reasonable to believe that a shift towards expenditure taxation would lead to significant increases in the private savings rate. I argue that the failure of traditional empirical approaches to isolate significant rate of return effects is a consequence of their failure to distinguish between transitory and permanent changes in the rate of return, and of other specification errors.

I. Theoretical Considerations

In a closed economy, it is not possible to imagine how the rate of return to savers could change without other relevant economic variables also changing. Thus, it is necessary to be clear about the nature of the shock causing the rate of return to change. Discussions of the "interest elasticity of savings" are apt to be misleading since the change in savings associated with any given change in the rate of return to savers will depend on what caused the rate of return to change. The analysis here focuses on the effects of tax policies which alter the rate of return available to savers. Any tax change will affect revenue collections and so must be associated with changes in either government spending, public borrowing, or other tax collections. The analysis here is based on a differential incidence approach, where it is assumed that spending and total revenue collections remain constant so that changes in capital income taxes are offset by adjustments to payroll or consumption taxes. The entire discussion is, therefore, about compensated effects. An effort is made to maintain this distinction in drawing implications from the empirical work in the discussion below.

The discussion here focuses on the "partial equilibrium effects" of a change in the rate of return. It is assumed that factor prices are unaffected by changes in the savings rate. Thus the analysis addresses the supply of savings schedule rather than the reduced-form relationship between tax changes and capital intensity. In the special cases of a small open economy, or a production func-

*Harvard University, Cambridge, MA 02138. This paper draws heavily on my working paper (1982).

tion with an infinite elasticity of substitution, the assumption of constant factor prices will be valid. Otherwise, it would be necessary to consider the aggregate production function in assessing the ultimate effect of a change in tax policy on private savings.

A. *Rule of Thumb Savings*

Economic theory needs to simplify reality enormously if anything tractable is to result. But it is important to acknowledge at the outset that no single analytic model can capture the complex motivations for any one individual's savings decisions, let alone the savings decisions of the entire population. The existence of substantial diversity in savings behavior creates a presumption in favor of a positive savings response to increases in the rate of return.[1]

Consider a population made up of "rule of thumb savers," each of whom saves a fixed fraction of his or her total disposable income regardless of the rate of return. The rule of thumb rate of saving varies across individuals; some are liquidity constrained and consume everything, others may have a quite high marginal propensity to save. Now imagine a reduction in the tax rate on capital income, financed by an equal revenue yield increase in labor income tax rate. Such a measure would, assuming some persistence in savings propensities, redistribute income from persons with low- to persons with high-savings propensities. As a consequence national savings would increase, even though no individual's savings incentive was affected. As time passes, the savings rate will rise further, as the share of total income going to persons with high-savings propensities increases.

B. *Life Cycle Savings*

Perhaps the dominant theoretical model used by economists in analyzing long-run questions relating to savings behavior is the life cycle hypothesis. In an earlier paper (1981), I argued that realistic formulations of the life cycle hypothesis implied a very substantial long-run response of capital accumulation to tax measures that change after-tax rates of return. The essential reason for the responsiveness of savings was the "human wealth" effect associated with changes in the after-tax rate of return. Increases in the after-tax rate of return reduce human wealth defined as the present value of individual labor income claims. This effect is absent in the two-period textbook formulations with all income received in the first period.

My conclusions about the high interest elasticity of savings have been challenged by Owen Evans (1983) and David Starrett (1982) who argue that they do not survive generalization of the model. Evans' principal point is that if one assumes a significantly negative time preference rate and a very low intertemporal elasticity of substitution, a relatively small interest elasticity of substitution will result. His point hardly seems warranted given that the elasticity is positive in every case he considers, and greater than .4 in most cases. Moreover, empirical evidence casts doubt on the relevance of the parameter values underlying Evans' low elasticity cases. Starrett shows that lower elasticities of savings can be generated using nonhomeothetic utility functions. However, both empirical evidence and theoretical considerations support the standard procedure of imposing homeotheticity. On balance, there remains reason to believe that life cycle saving is very likely to respond positively to after-tax rates of return, but the question is ultimately an impirical one.

C. *Bequest Savings*

My article with Laurence Kotlikoff (1981) suggests that bequests may account for a large fraction of national capital formation. The papers by Evans and Starrett discussed above argue that taking account of bequests makes it very plausible that the interest elasticity of savings is negative. The critical issue is how bequests are modelled. In my paper (1982), I establish the following results. As long as *any part of the population* is saving

[1] The argument here is developed rigorously in Laurence Seidman (1983) and my working paper.

for altruistic bequests, the long-run partial equilibrium elasticity of savings with respect to the rate of return will be infinite. Illustrative calculations suggest that it is likely to be very high in the short run as well. Thus taking account of bequests increases the predicted elasticity of savings.

Of course, alternative formulations of the bequest process are possible, although it seems hard to entirely rule out altruism. My own favorite is outlined in my paper with B. D. Bernheim and A. Shleifer (1983). It has implications similar to the standard life cycle model. Seidman incorporates bequests into my (1981) model by assuming that they generate utility directly for donors and get results qualitatively similar to mine.

The combination of the considerations discussed in this section suggest that the data should be approached with at least a mild presumption in favor of the hypothesis that savings respond positively to real after-tax rates of return. There are certainly internally consistent theoretical models which lead to a different conclusion, but their premises do not seem compelling.

II. Empirical Evidence

There have been many attempts to estimate the effects of changes in the rate of return on consumption and savings using Keynesian consumption functions. No consensus has emerged. I believe that there are fundamental conceptual problems which make it almost inconceivable that consumption function estimation can ever answer the questions of interest. Three difficulties seem paramount.

First, theory, particularly in the case of life cycle savers, suggests that the value of consumers' endowments is a function of the interest rate. Increases in the real after-tax interest rate reduce the value of human wealth, and may affect marketable wealth as well. These effects are not captured in standard formulations. When they are taken account of using a full macroeconomic model as in Franco Modigliani (1971), or a modified single equation consumption function as in my working paper, dramatic positive effects of increases in rates of return on savings result.

Second, the question of primary interest to persons concerned with tax policy is the response of savings to *permanent* changes in the real after-tax rates of return. The experiments provided by history came in the form of largely transitory changes in after-tax rates of return. Both theory and common sense suggest the response to temporary changes in rates of return should be much smaller than the response to permanent changes. This creates a strong presumption that simple extrapolation of the historical experience will lead to very substantial underestimates of the response of savings to permanent changes in the rate of return. This presumption is magnified by the very high noise-signal ratio in any attempted estimates of the real after-tax rate of return over a long horizon.

Third, there is the standard set of difficulties associated with any Keynesian consumption function. Almost all the right-hand side variables are probably endogenous. There is no satisfactory way of meeting the Lucas critique in modelling expected future labor income. No variables are included which address theoretically relevant issues such as the age structure of the population, or expected retirement ages. In an important recent study, Alan Auerbach and Kotlikoff (1981) illustrate the behavior of an economy in which the life cycle hypothesis holds exactly, and then fit standard consumption functions. The results indicate that parameter estimates are extremely sensitive to the choice of sample period, and that estimated parameters do not provide a useful guide to the effect of policy interventions.

What then can be done? I have suggested the futility of standard consumption function estimation for answering questions relating to long-run tax policies. Recent work by Lars Peter Hansen and Kenneth Singleton (1983), as well as others, suggests an alternative approach.[2] In general, it is possi-

[2] A survey of this burgeoning literature may be found in my article with Gregory Mankiw and Julio Rotemberg (1984). The method of estimation described here is frequently labelled the Euler equation approach.

ble to estimate the parameters of the utility function driving consumers' behavior, even where it is impossible to estimate any kind of structural consumption function. Essentially identification comes from the requirement that consumers satisfy certain first order conditions for utility maximization. This can be done using data on individual consumers as in David Runkle (1983) and Matthew Shapiro (1984), or with aggregation assumptions on aggregate data. Allowance can be made for the possibility that some consumers are liquidity constrained. Once utility functions have been directly estimated, simulation exercises of the sort performed in my 1981 paper can be used to estimate the effects of tax reforms. Of course, much more complex analyses taking account of individual diversity, and adding realistic information on wage earnings profiles should be possible.

At this point, the results of such elaborate simulation exercises cannot be predicted. However available evidence tends to suggest that savings are likely to be interest elastic. I find in the more reliable estimates in my working paper values of the intertemporal elasticity of substitution which cluster at the high end of the range Evans and I considered. Similar estimates are found using micro data by Shapiro, and by Hansen-Singleton. Where investigators find low estimates of intertemporal elasticity of substitution, it is usually because of the difficulty in modelling *ex ante* rates of return on corporate stock. It is also noteworthy that if proper allowance is made for trend growth in the economy, estimated time preference rates are positive, reinforcing the positive effects of higher rates of return on savings. Future research, particularly using micro data, will help to refine these conclusions and enhance our understanding of savings behavior.

III. Policy Implications

The U.S. economy now appears to be plagued by large structural budget deficits which appear likely to continue for the remainder of the decade unless major policy actions are taken. Private savings rates as measured in the National Income and Product Accounts do not appear to have increased along with the budget deficits. Indeed, many observers have expressed surprise that given the tax measures enacted in 1981, and the subsequent run upward in real interest rates, savings rates have not increased sharply. Some go as far as to call this a serious blow to supply-side economics.

Several observations should help to put this discussion in perspective. Unless savings are extraordinarily elastic with respect to rates of return, reductions in taxes will reduce the total supply of savings. Reduced public savings will not be offset by increases in private savings. The hope of those who advocated savings incentives was that in the long run, the revenue effects of these measures would be offset by reductions in spending or increases in other taxes. There is no serious case that permanent public dissavings to finance incentives is a viable strategy for raising national savings.

Does the stability of the private savings rate over the last several years constitute evidence against the view that savings respond positively to rate of return incentives? Probably the sample period is too short to permit conclusive judgements. Many other things happened over the last several years. For example, if the accrued gains to households on common stock are treated as part of income, the private savings rate was close to 20 percent over the last eighteen months. At the same time that wealth was rising rapidly, households were suffering through a severe temporary recession, tending to put further downward pressure on savings rates. A final factor working to make the private savings rate appear artificially low in recent years has been the erosion of inflation, which has led to unmeasured increases in real disposable income and savings.

These factors lead to the conclusion that the evidence is not in on the savings aspect of President Reagan's economic experiment. One of the few virtues of the macroeconomic turmoil we have suffered in recent years is that it has increased the power of our econometric experiments by raising the variance of most exogenous variables. Within a few years we should have made considerable progress towards resolving the uncertainties discussed in this paper.

REFERENCES

Auerbach, Alan J. and Kotlikoff, Laurence, "An Examination of Empirical Tests of Social Security and Savings, Working Paper No. 730, National Bureau of Economic Research, August 1981.

Bernheim, B. D., Shleifer, A. and Summers, L. H., "A Model of Manipulative Bequests," mimeo., 1983.

Boskin, Michael J., "Taxation, Saving, and the Rate of Interest," *Journal of Political Economy*, Part 2, April 1978, *86*, S3–27.

Evans, Owen J., "Tax Policy, the Interest Elasticity of Saving and Capital Accumulation," *American Economic Review*, June 1983, *73*, 398–410.

Hansen, Lars Peter and Singleton, Kenneth, "Stochastic Consumption, Risk Aversion, and the Intertemporal Behavior of Asset Returns," *Journal of Political Economy*, April 1983, *91*, 249–65.

Kotlikoff, Laurence and Summers, Lawrence H., "The Importance of Intergenerational Transfers in Aggregate Capital Accumulation," *Journal of Political Economy*, August 1981, *89*, 706–32.

Lucas, Robert E., Jr., "Econometric Policy Evaluation: A Critique," in Karl Brunner and Allen H. Metzler, eds., *The Phillips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl. 1976, 19–46.

Mankiw, Gregory, Rotemberg, Julio and Summers Lawrence H, "Intertemporal Substitution in Macroeconomics," *Quarterly Journal of Economics*, 1984 forthcoming.

Modigliani, Franco, "Monetary Policy and Consumption," *Consumer Spending and Monetary Policy: The Linkages*, Federal Reserve Bank of Boston, June 1971, 9–84.

Runkle, David, "Testing for Liquidity Constraints using Euler Equation Methods," mimeo., Brown University, 1983.

Seidman, Laurence S., "Taxes in a Life Cycle Growth Model with Bequests and Inheritances," *American Economic Review*, June 1983, *73*, 437–42.

______ **and Maurer, Stephen B.,** "Taxes and Capital Intensity in a Two-Class Disposable Income Growth Model," *Journal of Public Economics*, November 1982, *19*, 243–59.

Shapiro, Matthew, "The Permanent Income Hypothesis and the Real Rate: Some-Evidence from Panel Data," *Economics Letters*, No. 1, 1984, *14*, 93–100.

Starrett, David, "The Interest Elasticity of Savings," mimeo., Stanford University, 1982.

Summers, Lawrence H., "Capital Taxation and Accumulation in a Life Cycle Growth Model," *American Economic Review*, September 1981, *71*, 533–44.

______, "Tax Policy, the Rate of Return and Savings," Working Paper No. 995, National Bureau of Economic Research, September 1982.

GENDER ROLES IN ECONOMIC LIFE

Womenyouthandminorities and the Case of the Missing Productivity

By SHIRLEY P. BURGGRAF*

Women, youth, and minorities are frequently cited as an almost one-word explanation for declining productivity growth in the U.S. labor force as measured at the macro level. While some economists have proposed theories to account for productivity slowdown which emphasize such things as declining capital-labor ratio (J. R. Norsworthy, Michael Harper, and Kent Kunze, 1979); energy constraint (Edward Hudson and Dale Jorgenson, 1978); and macro instability (Richard Nelson, 1980; Michael Mohr, 1980), many analysts have cited the changing composition of the labor force as a significant contributing factor. (See, for example, Edward Denison, 1979; Norsworthy et al.; Frank Gollop and Jorgenson, 1980.) These assertions are tantamount to suggesting that admitting certain groups into the labor force necessarily lowers productivity and should not go unchallenged for several reasons. Inferring negative productivity coefficients to particular groups is likely to reinforce "statistical discrimination" whereby employers use perceived characteristics of groups as information shortcuts for making decisions about individuals (Edmund Phelps, 1972). It also has considerable implication for future U.S. productivity trends as the demographic composition of the labor force continues to change, and for formulating appropriate labor policies.

This paper examines labor market trends and productivity measurement methodology as it pertains to labor force composition, and concludes that whether women, youth, and minorities have contributed significantly to the decline in productivity growth, and are likely to do so in the future, depends on how productivity is defined and measured. Current definitions of productivity are essentially based on micro concepts. I argue that they may be misleading indicators of trends in economic efficiency at the macro level, especially under the conditions of social change and labor-supply growth which characterized the 1970's.

I. Empirical Context of the Issue

Historical trends in labor-force size, composition, and productivity are indicated in Tables 1 and 2 and can be summarized as follows:

1) The labor force has been the recipient of large numbers of new entrants, especially since the mid-1960's when an unprecedented decline in productivity growth occurred. While the labor force was increasing by some 32 million workers between 1966 and 1980, average annual productivity increases fell from 3.37 percent per annum to .93 percent per annum.

2) Youth, measured by percentage of work force under age 35, increased at an increasing rate until the early 1970's, continued to increase through the late 1970's, and is projected to decrease in the 1980's. In contrast with the female and minority components, youth will soon be a declining share of the labor force.

3) Women have accounted for an increasing share of the labor force for some time, and the trend is projected to continue at least through the early 1990's. Market work has become generally more important for females, and two-earner families and

*Department of Economics, Florida A & M University, Tallahassee, Florida 32307. Gratefully acknowledged are helpful suggestions and comments from Carol Jusenius, Joan Haworth, and Andrea Beller.

TABLE 1—CHANGES IN LABOR FORCE SIZE AND PRODUCTIVITY

Years	Increments[a]	Average Annual[b] Change in Productivity
1956–60	4,605	2.47
1961–65	4,817	3.37
1966–70	8,326	2.00
1971–75	11,004	1.87
1976–80	13,165	.93
1981–85[c]	8,045	
1986–90	7,390	
1991–95	5,167	

[a] *Source: Training and Employment Report of the President,* 1982, pp. 151–155.

[b] Shown in percent. *Source:* Department of Labor, BLS.

[c] *Source:* Medium-range Projections from *BLS Bulletin,* "Economic Projections to 1990," No. 2121.

TABLE 2—CHANGES IN LABOR FORCE COMPOSITION

	Changes in Percent of Labor Force that is:[a]		
Years	Under 35	Female	Nonwhite
1956–60	−.93	1.77	.40
1961–65	.92	1.82	.10
1966–70	3.74	2.91	−.04
1971–75	5.24	1.86	.58
1976–80	2.48	2.57	.81
1981–85	−1.10	2.15	.29
1986–90	−2.80	1.48	.73
1991–95	−4.09	.74	.82

[a] See sources given in Table 1.

female-headed households are an increasing factor in the labor force.

4) Minority workers have also been a steadily increasing component of the work force, except for one period, and are expected to continue to increase. However, the order of magnitude for minorities is much smaller than for youth or women.

The diversity of timing and magnitude in these trends makes them unlikely candidates for aggregation in explaining labor force productivity. They are, however, routinely grouped together, apparently as a consequence of the particular way productivity is defined and measured.

II. Methodology of Productivity Measurement

Productivity analysis usually begins with a production function or a growth accounting equation which under certain conditions can amount to the same thing (Mohr). Following Mohr, if there is a production relationship such as

$$(1) \qquad Y=f(Z,t),$$

where Z represents n input types, then the rate of output growth over time t can be measured as

$$(2) \quad (\Delta Y/\Delta t)/Y=\sum_i f_i Z_i/f\cdot(\Delta Z_i/\Delta t)/Z_i + (\Delta f/\Delta t)/f,$$

where f_i is the marginal product of input Z_i and $(\Delta f/\Delta t)/f$ is the productivity residual which is alternately termed the growth of technological progress or the rate of total factor productivity.

Assuming the input vector Z can be partitioned into aggregate input types such as $Y=f(K,L,E,M)$ where K,L,E,M represent capital, labor, energy, and materials; letting A represent the productivity residual; and assuming pure competition and constant returns to scale, equation (2) can be written

$$(3) \quad \dot{Y}/Y=P_K K/P_Y Y\cdot\dot{K}/K + P_L L/P_Y Y\cdot\dot{L}/L+P_E E/P_Y Y\cdot\dot{E}/E + P_M M/P_Y Y\cdot\dot{M}/M+\dot{A}/A$$

where $P_K K/P_y Y$, etc., are the cost shares of the macro inputs in the value of output. Equation (3) can be transformed into a measure of labor productivity growth ($P\dot{R}_L = \dot{Y}/Y - \dot{L}/L$) by subtracting $\dot{L}/L$ from both sides, which gives

$$(4) \quad P\dot{R}_L=P_K K/P_Y Y\cdot\dot{K}/K + [P_L L/P_Y Y-1]\dot{L}/L+P_E E/P_Y Y\cdot\dot{E}/E + P_M M/P_Y Y\cdot\dot{M}/M+\dot{A}/A.$$

Equation (4) is computed for industries at various levels of disaggregation, and national productivity is a weighted summation of equation (4) across industries.

In the context of equation (4), perceived changes in labor quality can be introduced as an adjustment to the input variable L or as a component of the residual A attributable to labor-quality effects. Denison and Gollop and Jorgenson are two typical approaches which demonstrate the alternatives of treating labor quality as an "A" effect or an "L" effect, respectively. Denison counts all hours worked as equal increments to L and then estimates effects of labor-quality changes on A. He specifically notes that while such disparate groups as nonfarm wage and salary workers, nonfarm self-employed workers, farm workers, and unpaid family workers can be viewed as homogeneous labor inputs "hours worked by persons in different age-sex groups do not represent the same input" (p. 2). His rationale is a simple application of the neoclassical tenet that people are paid less in the market place because they produce less. On the basis of their relatively lower market wages, he estimates that increases in youth and female components of the labor force decreased residual productivity growth by .17 percentage points per annum from 1948 to 1973 and by .25 points per annum between 1973 and 1976.

Gollop and Jorgenson employ the same market-based assumptions about relative worth of various groups of workers but propose to deal with "quality" variations in labor input directly by adjusting L rather than allowing quality effects to influence A. Gollop and Jorgenson adjust L for two sexes, eight age groups, five levels of education, two employment classifications, ten occupations, and fifty-one industries by weighting each category with its relative market wage to compute a Divisia index of labor input. Since female workers earn less than male workers, their labor input is discounted proportionately. Because a portion of their input doesn't exist for measurement purposes, the productivity residual represented by A in equation (4) with any given level of Y is made to appear larger by this procedure.

III. Measurement Anomalies

Productivity models like those outlined above contain some measurement anomalies which particularly relate to the age-race-sex issue. Issues which deserve attention include 1) an asymmetric treatment of capital and labor utilization; 2) labor market distortions such as wage discrimination, occupational segregation, segmented markets, or other forms of market distortion which may affect wage patterns as indicators of productivity but not be considered in the macro productivity measure; and 3) disparate treatment of particular groups and its impact on productivity measurement.

A. *Factor Utilization Asymmetry*

It is generally assumed that all undepreciated plant and equipment is utilized whereas idle labor is not. The justification most often cited is given by John Kendrick (also cited by Gollop-Jorgenson, p. 111):

> In contrast to the human population the entire living population of capital goods is available for productive use at all times and involves a per annum cost, regardless of degree of current output and income. The degree of capital utilization reflects the degree of efficiency of enterprises and the social economy generally. Hence, in converting capital stock in inputs, we do not adjust capital for changes in rates of capacity utilization, and thus these are reflected in changes in productivity ratios. [1973, p. 26]

By this rationale all existing capital is counted as capital input whereas only employed labor is counted as labor input. This procedure is consistent from a business cost perspective, but inconsistent from the standpoint of social cost and macro policy since unemployed labor is a waste of resources as much as unutilized capital. The cohort of new labor force entrants in the 1960's and 1970's could not have depressed productivity statistics if they had remained unemployed but inevitably did so by becoming employed. The bracketed term in equation (4), which measures the effect of increase in labor input on productivity growth, is always negative *regardless of the personal characteristics of the new entrants*. "Productivity" can, therefore, be a somewhat perverse economic indicator in times of labor surplus because of an inherent conflict between maximizing production and employment and maximizing labor productivity as it is currently measured (see my 1981 paper, pp. 56–61). A definition which reconciles movements in production with movements in productivity would be output/total labor available rather than output/labor employed. By defining labor productivity as output per available worker whether employed or not (as capital input is measured), production and productivity would be maximized at the same level of output (i.e., full employment).

B. *Wage/Productivity Distortions in the Labor Market*

The many controversial and unsettled questions concerning existence, extent, and form of labor market discrimination are not addressed here. Instead I focus on the implication of labor-market imperfections on productivity growth and its measurement. Obviously, if people are paid unequal wages for equal output, weighing labor inputs by market wages distorts "true" productivity estimates. Other kinds of distortions, however, may be less obvious.

One example of a labor market phenomenon that is probably distorting productivity measurement is a variation of the segmented market hypothesis. It is a conspicuous empirical fact that women and minorities tend to be concentrated in crowded, low-paying, deadend jobs (Ralph Smith, 1979, p. 43). Such jobs also frequently serve supporting roles for white male workers. One extreme example of a low-paying support job is that of the full-time housewife, counted as zero input in the labor market because she receives no wage, who contributes significantly to her husband's productivity. Measuring housewives' contributions to their husbands' productivity is difficult, but may perhaps be approximated by a marriage coefficient in earnings analyses (see Charles Haworth and Joan Haworth, 1975). In any event, no one would argue that the contribution is always zero. If some part of the productivity of workers in an unmeasured "support sector" of the economy is captured by workers in a measured sector, then movement of workers out of traditional support jobs would lower the measured productivity of workers in the measured sector.

It can be argued that women did not "go to work" in the 1970's—they changed jobs—from performing backup roles as homemakers for male wage earners to participating more directly in the paid labor force. Some employers who were previously getting two (a husband with a full-time wife) for the price of one are now just getting one employee. It may also be argued that employers sometimes get two for the price of one and one-half—the primary worker backed up by talented subordinates in support positions, such as the manager and his equally talented secretary. It should be no surprise that a white male worker who was formerly backed up by a full-time wife at home and an overqualified secretary at the office, educated by talented teachers with few career alternatives, and perhaps even served by a maid, handyman, janitor, or cleaning lady, could lose productivity if he loses part of a traditional support system when women and minorities "go to work."

C. *Disparate Treatment*

Whether changes in labor force composition are treated as residual effects or input effects in the productivity equation hardly matters because both approaches essentially define away the productivity problem as it applies to women, youth, and minorities. If the major concern with productivity in the first place is for what it represents for real wages and standard of living, there is something circular about weighting labor inputs with market wages. While productivity growth as measured by A in equation (4) is being discussed as a major national concern, designating a component of A as being that of women and youth when no other groups are so designated seems to be a way of saying that their productivity (and therefore their lower wages and lower standard of living) are of less concern. More explicitly, subtracting an age-sex component from L in equation (4) essentially says that those labor inputs in some sense don't exist so there can't be any missing productivity to cause concern. For young workers who presumably have better things ahead, such a perspective may be socially justified; but for female and minority workers of varying ages it is a different matter.

IV. Conclusions

Women, youth, and minorities are not a one-word factor in the productivity puzzle. Their historical trends are different and they have played different roles in the labor market. Inexperience is sometimes cited as a reason for grouping them together, but it cannot be assumed that inexperience is evenly shared by such groups. The female percentage of the labor force is undoubtedly rising partly because women are staying on the job rather than leaving when they start families, thereby adding to the experience of the labor force rather than subtracting. In any event, youth and inexperience are conditions which pass with time, whereas the condition of being a woman or minority is permanent; so they have very different implications for long-term productivity trends. What women, youth, and minorities really have in common and the apparent reason why they are grouped together for productivity measurement is their low wages; but since wages are presumably a result of productivity not a cause, wages cannot be an explanatory factor in productivity analysis—they can only be used to redefine the problem.

The most notable fact about the new labor force entrants of the 1970's was not their age, race, or sex, but their numbers—some 24 million of them. Other than an extraordinary rate of capital accumulation or technological breakthrough, the only way to have kept "productivity" up would have been to keep a lot of potential workers unemployed.

Since it is difficult for women or minorities to bring the same network of support workers to higher level jobs that white males have had (because women and minorities have been the support network), their measured productivity in some cases will inevitably be lower than historical norms as will that of white male workers who are losing part of their support system in the process of social change.

The productivity model has always been incomplete at the macro level because of its inconsistent definitions of capital and labor inputs, its omission of the household sector, and various anomalies in the labor markets. Changes which occurred in gender roles in economic life in the 1970's have undoubtedly served to exacerbate the problems. If the variety of measurement problems and anomalies identified here in connection with labor force composition is some indication, productivity at the macro level may prove to be an elusive concept.

REFERENCES

Burggraf, Shirley P., "The Job-Generation/Size-of-Firm Issue: Critique and Synthesis," Economic Development Administration, Washington 1981.

Denison, Edward F., "Explanations of Declining Productivity Growth," *Brookings General Series*, Reprint 354, Washington: The Brookings Institution, 1979.

Gollop, F. M. and Jorgenson, D. W., "U.S. Productivity Analysis," in J. W. Kendrick and B. N. Vaccara, eds., *New Developments in Productivity Measurement and Analysis*, New York: National Bureau of Economic Research, 1980.

Haworth, Charles and Haworth, Joan, "Progress or Regression?—Earnings Differentials During the 60's," *Proceedings of the American Statistical Association*, 1975.

Hudson, E. A. and Jorgenson, D. W., "The Economic Impact of Policies to Reduce U.S. Energy Growth," in *Resources and Energy*, Amsterdam: North Holland, 1978.

Kendrick, John W., *Postwar Productivity Trends in the United States, 1948–1969*, New York: National Bureau of Economic Research, 1973.

Mohr, Michael F., "Concepts and Measurement of Productivity," paper presented to the American Productivity Conference, Houston, Texas, 1980.

Nelson, Richard, "Technical Advances and Productivity Growth: Retrospect, Prospect, and Policy Issues," in *Western Economies in Transition: Structural Change and Adjustment Policies in Industrial Countries*, Hudson Institute Studies on the Prospects of Mankind, Boulder: Westbrook Press, 1980.

Norsworthy, J. R., Harper, Michael J. and Kunze, Kent, "The Slowdown in Productivity Growth: Analysis of Contributing Factors," *Brookings Papers on Economic Activity*, 2:1979, 387–421.

Phelps, Edmund, "The Statistical Theory of Racism and Sexism," *American Economic Review*, September 1972, *62*, 659–61.

Smith, Ralph E., *The Subtle Revolution*, Washington: Urban Institute, 1979.

Work Characteristics and the Male-Female Earnings Gap

By MARIANNE A. FERBER AND JOE L. SPAETH*

Much research has been done to explain the continued disparity in men's and women's earnings, and progress has been made in demonstrating factors associated with this gap. Nonetheless, disagreement remains concerning the appropriate variables to be used and the statistical techniques to be employed. This paper focuses on the former question.

Section I consists of a brief review of some of the best known studies in the field. Section II suggests a different approach for shedding light on some unresolved issues, and describes the data set used. Section III contains the empirical analysis and discussion of the results. Tentative conclusions and suggestions for further work are offered in Section IV.

I. Findings of Previous Studies: Review and Comparisons

A recent comprehensive survey by Donald Treiman and Heidi Hartmann (1981, pp. 19–41) reviewed studies that approached the earnings gap from the point of view of human capital theory. Estimates of the proportion of variance explained by differences in qualifications of workers varied from 0 to 44 percent. The wide range and the small proportion explained by most estimates both deserve attention.

The seven studies reviewed differ considerably in terms of the data used, the explanatory variables included, the statistical methods employed, and the precise definitions of the dependent variables. A careful examination, however, shows that it is the use of both a measure of proxy or actual labor market experience and of occupational training, that is associated with the two highest estimates, 44 percent for Mary Corcoran and Greg Duncan (1979) and 41 percent for Jacob Mincer and Solomon Polachek (1974). Ronald Oaxaca (1973), who uses a proxy for experience gets an estimate of 20 percent. The others use neither and obtain estimates ranging from zero for Alan Blinder (1973), to 18 percent for Isabel Sawhill (1973). None explain as much as half of the earnings gap.

All the studies cited above rely merely on characteristics of individual workers and do not take into account occupational differences between men and women. It has been argued that virtually all of the earnings gap could be accounted for by differences in occupational distribution if we had a sufficiently detailed breakdown of categories (Victor Fuchs, 1971). Among nine studies that do take occupation into account in one way or another, none come close to fulfilling this claim; they do, however, have considerably more explanatory power than the ones which omit occupations. The portion of the gap explained ranges from 15 percent (Fuchs) to 43 percent (Treiman and Kermit Terrell, 1975). Each of the studies relies on a somewhat different assortment of personal characteristics and different data sets, but none use a really detailed classification of job categories.

Much of the remaining difference may be attributed to sex segregation by firm. Donald McNulty (1967), John Buckley (1977), and Francine Blau (1975) all documented that women tend to be employed in firms that pay less than those primarily employing men. When men's and women's earnings are compared within the same establishment, as well as within detailed job classifications, the earnings differential does almost disappear (Marina Whitman, 1973).

The problem with this approach is that for the most part men and women are not in the same detailed job classifications, and not in the same establishments, so that rather than "explaining" why women are paid less, we merely shift to the question why female oc-

*Department of Economics, and Department of Sociology, respectively, University of Illinois, Urbana, IL 61801.

cupations and firms employing females pay less than male occupations and firms employing males.[1] Furthermore, this approach does not provide any help in determining when workers in different jobs and different firms do substantially similar work, an issue that is becoming of considerable practical importance.

II. Data on Work Characteristics

One way to avoid the problems of using detailed job categories is to investigate specific work characteristics that are common to a wide variety of occupations, both male and female. We have been investigating a group of characteristics that involves control over job-related resources, including the work of others, organizational policies, and monetary resources. The hypothesis to be investigated here is that these characteristics, along with structural determinants, have important wage effects over and above the effects of human capital.

Recent work, much of it by sociologists, has given some support to this view. Charles Halaby (1977), Martha Hill and James Morgan (1979), Wendy Wolf and Neil Fligstein (1977), and Eric Wright and Luca Perone (1977) all have examined attributes of particular jobs, especially authority of the position, and concluded they are useful predictors of earnings. The variables these researchers used were, however, rather general, and did not permit a realistic representation of the multiple hierarchical levels that exist in large organizations. None investigated authority over financial resources.

The present study uses data collected for this purpose and provides considerably more detailed information both on the structural setting where the jobs are performed and the specific work characteristics of individual jobs. This enables us to determine the importance of these factors in influencing earnings and incidentally to learn whether men and women are comparably rewarded for comparable work, since none of the work characteristics or the structural features are unique to men or women.

The data were collected as part of a practicum in survey research methods. Students trained by professional supervisors carried out telephone interviews in the spring of 1982 with 557 persons living in the state of Illinois who were employed at least 20 hours a week on a single job. Owing to the prevalence of nonlisted telephones, random digit dialing was used in the Chicago *SMSA*. A systematic sample was drawn from telephone directories for the area outside Chicago. In order to provide sufficient cases for separate analysis by sex, a selection procedure was used to increase the number of women. It yielded a sample of 44 percent women and 56 percent men.

The usual questions were asked about work history and current job as well as classifying characteristics of the individual. In addition, detailed information designed to measure quite specific work characteristics was obtained, including supervisory authority, discretion of respondents and their subordinates, scope of the respondent's policymaking activities, activities involved in spanning of the organization's boundaries, and control over monetary resources.

III. Analysis

In order to determine the extent to which some of the work characteristics that would be expected to be relevant help to explain men's and women's earnings, we added these to an otherwise standard regression with ln of total earnings as the dependent variable. Different segments of work experience, number of hours and weeks worked, number of years of schooling, and marital status are the independent variables most usually included. Whether the respondent works in a "core" industry,[2] and size of the firm, both in terms

[1] Polachek (1976) has attempted to explain existing occupational segregation in human capital terms, but his conclusions have been challenged (Paula England, 1982), and other researchers have found evidence that women face institutional barriers in their occupational choices (Duncan and Saul Hoffman, 1979; Carol Schreiber, 1979).

[2] E. M. Beck et al. (1978) devised a classification for core and peripheral sectors, where industries were assigned to the periphery because of their small firm size,

of ln of number of employees and of geographic boundaries, were added as a test of the dual labor market theory. Of the work characteristics we finally entered one measure of supervisory authority over other units, and an index of monetary control, to test the extent to which power over resources is rewarded. Last, we added sex of the respondent's supervisor, to test the hypothesis that having a male boss is rewarded. (The table and a description of the variables are available upon request.)[3]

The coefficients for the traditional variables are consistent with the predictions of human capital theory. For men, all three segments of experience are significant at the 1 percent level and the size of the coefficients is about the same for each segment. For women, only years on current job are significant. Hours and weeks worked are significant for both men and women, at the 5 percent level or better. Marital status is not significant for either group, nor is schooling for women. Both may, however, have indirect effects via influence on experience and type of work.

Our findings with respect to the variables related to employers are consistent with the dual labor market theory. Men but not women are significantly rewarded for being in a core industry, suggesting that women often have peripheral jobs, regardless of the industrial sector in which they are located. Similarly, only men benefit from working in a firm that has broad geographic coverage, perhaps because they are more willing and able to take advantage of the job opportunities available in such organizations. Size of firm, in terms of number of employees, on the other hand, significantly benefits only women. Equal opportunity legislation, often only applicable to and probably more rigidly enforced in larger firms, may be the explanation.

Of particular interest are the results with respect to the specific job characteristics. Sex of supervisor is significant at the 1 percent level for women, and increases their earnings by almost 25 percent. For men it is only significant at the 10 percent level, but this is not surprising given that 94 percent of them have male supervisors. Thus it would appear that there is considerable advantage to having a male boss, or, to put it differently, a substantial penalty for having a female one.

Monetary control turns out to be significant at the 5 percent level or better for both men and women, and substantially increases the earnings of those who have such authority. While the coefficient is slightly higher for women, they have significantly less control, and the slightly lower percentage increase represents a far higher dollar amount for men. In a regression identical with the one used here, but with dollar earnings as the dependent variable, the coefficient for men was more than twice as large as for women.

We also calculated how much of the earnings gap is caused by the differences in the mean values of variables for men and women, as opposed to differences in the reward structure. This was done using the mean values for women, multiplied by the male coefficients, as is the usual practice. The results showed that 56.6 percent is caused by the differences in the variables. Only studies using detailed occupational categories have equalled or exceeded these results.

IV. Conclusions

Single minded proponents of one particular explanation of earnings differentials will find little joy in our results. They do not support the importance of one factor to the exclusion of all others, but rather add evidence that many play a significant part.

As indicated earlier, our study confirms that human capital is rewarded for both men and women. That the rewards relative to some variables are unequal may to some extent be caused by differences in the quality

seasonal and other variations in product supply and demand, labor intensity, weak unionization, and low assets.

[3] The R^2s obtained of .497 for men and .389 for women compare quite favorably with those in most previous studies, even though we have no information on the respondent's vocational training, state of health, or timing of work interruptions.

of the capital accumulated, as for instance in years of schooling and experience. We also find, however, that rewards are influenced by the setting in which the human capital is employed, suggesting that dual labor market theory also makes a contribution. In addition we uncovered evidence that two job characteristics not generally considered in previous research have considerable independent influence, not readily explained by either of the established theories.

The findings with respect to sex of supervisor and monetary authority have interesting implications, all the more so because workers tend to be assigned to supervisors, and control over financial resources similarly tends to be assigned to them. The fact that a worker earns more when the immediate supervisor is male is consistent with the hypothesis that his sex confers higher status on an otherwise comparable job.[4] To the extent that there is discrimination in this respect,[5] the general presumption that workers tend to be rewarded according to their productivity alone is undermined.

The finding that control over money, a highly rewarded and quite specific characteristic is associated with higher dollar rewards for men than women, also points toward the existence of discrimination. If additional research discovers more such work characteristics that can be precisely measured, it will help us to move toward an operational definition of what is "work of comparable worth." This is a necessary first step if the present rule of equal pay for equal work, which has done so little to close the earnings gap between men and women, is to be broadened to the potentially far more effective equal pay for comparable work.

[4] There is no evidence to suggest that men are better supervisors. On the contrary, two studies showed that workers were as satisfied with female as with male supervisors. Both also found that people who had worked with such women had far more positive attitudes toward female supervisors (Ferber, Joan Huber, and Glenna Spitze, 1979; Hubert Field and Barbara Caldwell, 1979).

[5] Sex of supervisor may to some extent pick up the effect of being in a "female" occupation. Our data do not enable us to distinguish between these hypotheses.

REFERENCES

Beck, E. M., Horan, Patrick M. and Tolbert, Charles M. II, "Stratification in a Dual Economy," *American Sociological Review*, October 1978, *43*, 704-20.

Blau, Francine, D., "Sex Segregation of Workers by Enterprise in Clerical Occupations," in R. C. Edwards et al., eds., *Labor Market Segmentation*, Lexington: Lexington Books, 1975.

Blinder, Alan S., "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, Fall 1973, *8*, 436–55.

Buckley, John E., "Pay Differences Between Men and Women in the Same Job," *Monthly Labor Review*, November 1977, *94*, 36–39.

Corcoran, Mary and Duncan, Greg J., "Work History, Labor Force Attachment, and Earnings Differences Between Races and Sexes," *Journal of Human Resources*, Winter 1979, *14*, 3–20.

Duncan, Greg J., and Hoffman, Saul, "On-the-Job Training and Earnings: Differences by Race and Sex," *Review of Economics and Statistics*, November 1979, *61*, 594–603.

England, Paula, "The Failure of Human Capital Theory to Explain Occupational Sex Segregation," *Journal of Human Resources*, Summer 1982, *17*, 358–70.

Field, Hubert S. and Caldwell, Barbara E., "Sex of Supervisor and Sex of Subordinate," *Psychology of Women Quarterly*, Summer 1979, *3*, 391–99.

Ferber, Marianne A., Huber, Joan A. and Spitze, Glenna, "Preference for Men as Bosses and Professionals," *Social Forces*, December 1979, *58*, 466–76.

Fuchs, Victor, "Differences in Hourly Earnings Between Men and Women," *Monthly Labor Review*, May 1971, *94*, 9–15.

Halaby, Charles N., "Job-Specific Sex Differences in Organizational Reward Attainment: Wage Discrimination vs. Rank Discrimination," Discussion Paper No. 469–77, Institute for Research on Poverty, 1977.

Hill, Martha and Morgan, James N., "Dimensions of Occupation," in Greg J. Duncan and James N. Morgan, eds., *Five Thousand*

American Families—Patterns of Economic Progress, Ann Arbor: ISR, University of Michigan, 1979, 293–334.

McNulty, Donald J., "Differences in Pay between Men and Women Workers," *Monthly Labor Review*, December 1967, *90*, 40–43.

Mincer, Jacob and Polachek, Solomon W., "Family Investments in Human Capital: Earnings of Women," *Journal of Political Economy*, March/April 1974, *82*, S76–108.

Oaxaca, Ronald, "Male-Female Wage Differentials in Urban Labor Markets," *International Economic Review*, October 1973, *14*, 693–709.

Polachek, Solomon W., "Occupational Segregation Among Women: A Human Capital Approach," U.S. Department of Labor Report No. ASPER/PUR-75/1909A, April 1976.

Sawhill, Isabel V., "The Economics of Discrimination Against Women: Some New Findings," *Journal of Human Resources*, Summer 1973, *8*, 383–95.

Schreiber, Carol T., *Changing Places: Men and Women in Transitional Occupations*, Cambridge: MIT Press, 1979.

Treiman, Donald J. and Hartmann, Heidi I., *Women, Work and Wages: Equal Pay for Jobs of Equal Value*, Washington: National Academy Press, 1981.

______ **and Terrell, Kermit**, "Sex and the Process of Status Attainment: A Comparison of Working Women and Men," *American Sociological Review*, April 1975, *40*, 174–200.

Whitman, Marina v.N., *Testimony* before U.S. Congress, Joint Economic Committee, Part 1, Washington 1973.

Wolf, Wendy C. and Fligstein, Neil D., "Sexual Stratification: Differences in Power in the Work Setting," Working Paper No. 77–19, Center for Demography and Ecology, University of Wisconsin, 1977.

Wright, Eric O. and Perone, Luca, "Marxist Class Categories and Income Inequality," *American Sociological Review*, February 1977, *42*, 32–55.

An Economic Model of Asset Division in the Dissolution of Marriage

By CAROL C. FETHKE*

Recent statistics indicate that more than one-third of all new marriages will end in divorce. This evidence suggests that even the most "happily married" couples may be wise to view their lifetime choices within a framework that recognizes that periodically each selects one of two strategies: married or not married. When both select the married strategy, the couple remain married. If either party (or both) elect the not-married strategy, the outcome will be divorce. Election of the not-married strategy thus creates a two-period world in which each party is married in the first period and divorced in the second.

When a marriage dissolves, the couple divides all marital property either by mutual consent or according to the division rules imposed upon them by the state in which they reside. The law separately defines both marital property and the formula used to divide the property. This paper focuses on the interaction of the two variables, specifically, the effect of the state's division rule on the savings-consumption decisions of a divorcing couple.[1] Savings are of interest because they represent the couple's marital assets; the division rule is important because the amount each party receives at divorce affects the postmarriage economic well-being of each. The analysis can improve our understanding of the economic behavior of couples and will provide insight into the effect of divorce law on family savings patterns.

Since law views divorcing spouses as adversaries, the analysis of marital savings and the resulting property division is carried out in a noncooperative game framework in which couples facing divorce each protect their self-interest by maximizing separate lifetime utility functions.[2] Three noncooperative games are discussed: Cournot, Stackelberg, and Nash bargaining.

I. The Model and Some Assumptions

Let us consider a couple married in the first period and divorced at the beginning of the second. Each party maximizes a separate two-period lifetime utility function whose only arguments are their own consumption in the two periods. Specifically,

$$(1)\quad \operatorname{Max} U^i = \left(c_1^i c_2^i\right)^{1/2} \quad U' > 0, \quad U'' < 0,$$

$$\text{subject to} \qquad \sum c_1^i \leq \sum y_1^i,$$

$$c_2^h = y_2^h + (1-\theta)S_1^m, \quad c_2^w = y_2^w + \theta S_1^m,$$

where $i = h =$ husband, $w =$ wife, $j = 1 =$ marriage, $2 =$ divorce, and U^i are the utilities of each spouse; c_j^i is the consumption of spouse i in period j; y_j^j is the income stream of i in j. Marital savings, S_1^m, is defined as $S_1^m = \Sigma y_1^i - c_1^h - c_1^w$. The wife's share of marital assets as prescribed by law is θ; $(1-\theta)$ is the husband's share, $0 \leq \theta \leq 1$.

Since the question of interest is how incomes and property division laws affect marital savings patterns, the utility function selected is a geometric mean which maximizes utility when consumption in the two

*Department of Home Economics, 127 Macbride Hall, University of Iowa, Iowa City, IA 52242. I thank Guillermo Owen for his assistance, and Andrew Daughety, Robert Forsythe, and Gary Fethke for their helpful comments.

[1] A happily married couple who always elects the married strategies will have no second period and therefore no reason to save. Divorcing couples save to transfer marital income to the second period.

[2] Typically, single individuals are described as maximizing their lifetime utilities; couples are described as maximizing a family utility function or separate utility functions that include the well-being of their spouse or family members as one of the arguments. The latter frameworks do not lend themselves readily to couples who are facing divorce, thus the selection of separate lifetime utilities for this model.

periods is equal. In this way no preference is given to consumption in one time period over the other. For a similar reason both the discount rate and rate of return on savings are omitted (or assumed to be equal to one). Later, removal of these restrictions is discussed.

The first constraint states that no savings will take place if marital consumption equals the marital income stream; if the inequality holds, savings occur. Note that this constraint also prevents savings from being negative, that is, no borrowing from the divorce period is permitted in the model.

The second and third constraints specify that consumption for each spouse after the divorce depends upon their own divorce-period income stream, y_2^i, plus their share of marital assets.[3] Marital assets are the savings that result from pooling all marital income not spent by one of the parties during the marriage. This single asset, S_1^m, is a realistic way of describing the family's portfolio of home, pensions, securities, etc., which will be subject to division. It is the pooling of income and any resulting savings that distinguishes the analysis of a married couple from that of two separate individuals. Throughout, no assumption is made, a priori, about the relative levels of the income streams of either spouse in either period.

Two final assumptions have been made to simplify the structure of the problem. No provision has been made for a variable to represent a premarital asset base. Since the legal definition of marital property subject to division excludes premarital assets, this omission is equivalent to reducing c_j^i by whatever each party owns separately. Premarital assets could most easily be incorporated by redefining Σy_1^i to include such assets and the income from them and to define y_2^i to include premarital assets of each party not consumed during the marriage. The second simplifying assumption has been to ignore all transactions costs of divorce. If divorce costs are fixed, they can easily be incorporated into the model by redefining marital property as marital savings minus divorce costs. If, however, costs are variable, perhaps a function of bargaining time for example, they would have to be specifically incorporated into the model and the results discussed here would have to be modified.

Given this basic structure, three noncooperative games are considered. In the Cournot framework, each partner selects a level of c_1^i to maximize equation (1) believing that their spouse will not be influenced by their choice; each takes the other's level of consumption as invariant. A consistent equilibrium can result only if their reaction curves intersect.

In the Stackelberg model, the husband is assumed to be a "leader" and the wife a "follower." She selects first-period consumption assuming her husband's behavior is fixed and unresponsive to choices she makes; she maximizes equation (1). He, on the other hand, maximizes his lifetime utility by modifying equation (1) to take into account his wife's reaction function.

In the Nash-bargaining problem, the partners maximize their respective lifetime utilities by dividing the total utility available from income in periods one and two. They can achieve this division either by agreeing to a Pareto optimal solution, by carrying out their threats, or by negotiating a bargained solution.

It is appropriate to comment on these three games. The fixed behavior of the Cournot framework is probably not realistic, although it is a familiar game and therefore provides readily recognizable results. The Stackelberg framework (leader/follower) is more appealing, but suffers from the problems of inconsistency if both spouses try to lead. The Nash-bargaining model is perhaps the most realistic. It provides for Pareto optimal solutions if the couple agree to a division. If not, lawyers can convey the threat points, and assist in bargaining for mutual gains in utility.

[3] The second-period income stream of each can be thought of as the couple's prediction of divorce-period incomes. It is not necessary to assume that each party knows what the true future values of income will be, only that they agree on the predictions when making their choices. Obviously, an over- or understatement of the values of y_2^i (when compared to true values of y_2^i) will lead to choices of c_1^i, which may not actually optimize U^i. The problems of incorrectly estimating y_2^i or of assuming different values for y_2^i than those used by a spouse are not discussed.

II. Solutions to Games Couples May Play in Divorce

In all three models, the choice variables are consumption during the marriage. In the Cournot game, consumption in period ($t=1,2$) by husband, c_t^h, and wife, c_t^w, have the following equilibrium values:

$$c_1^h=\frac{1}{3}\left[\sum y_1^i+\frac{2y_2^h}{1-\theta}-\frac{y_2^w}{\theta}\right]$$

$$c_2^h=\frac{1}{3}(1-\theta)\left[\sum y_1^i-\frac{y_2^h}{1-\theta}-\frac{y_2^w}{\theta}\right]$$

$$c_1^w=\frac{1}{3}\left[\sum y_1^i-\frac{y_2^h}{1-\theta}+\frac{2y_2^w}{\theta}\right]$$

$$c_2^w=\frac{1}{3}(\theta)\left[\sum y_1^i-\frac{y_2^h}{1-\theta}-\frac{y_2^w}{\theta}\right].$$

For savings to be positive, as assumed, it can be shown that θ must satisfy the constraint

$$(2)\qquad y_2^h/(1-\theta)+y_2^w/\theta\leq\sum y_1^i.$$

The resulting condition requires that the total marriage-period income stream exceed the sum of the separate divorce incomes where each y_2^i is weighted by the inverse of the spouse's own division share. As θ approaches zero or one, this condition is unlikely to be met. In fact, it appears to be remarkably limiting even if θ is near 1/2. The condition might be met by a couple who were married for the majority of their adult lives so that the anticipated divorce-period income stream was small because the second period was short. However, most divorcing couples are likely to be both young and not long married. Given a young individual's typical earnings profile, it is unlikely that most divorcing couple's marital incomes will exceed their divorce incomes by enough to meet the conditions the model demands. Thus, one would expect most marriages ending in divorce to have no marital property to divide.

When savings are positive in the Cournot game, the above results suggest that each party's marital consumption level depends on the interrelationship between second-period incomes and θ. During the marriage they will have equal consumption only when $y_2^h/(1-\theta)=y_2^w/\theta$. A couple will save more (spend less) as the difference between $\Sigma y_1^i-(y_2^h/(1-\theta)+y_2^w/\theta)$ is greater. For any fixed level of postmarriage incomes, marital savings will be greatest when $y_2^h=y_2^w$ and $\theta=1/2$. Given $\Sigma y_2^i=k$ and $y_2^h=y_2^w$, as θ falls below 1/2, the husband's marital consumption will fall, the wife's increase. As θ moves away from 1/2 in either direction, marital savings will fall because the spouse receiving the smaller share of property will have less incentive to save.

In general, $dS_1^m/d\theta=0$ implies $(y_2^w/y_2^h)^{1/2}=\theta/(1-\theta)$. Given $\theta<1/2$ and a given level of postmarriage income, a couple will save more when $y_2^h>y_2^w$. As the difference between the husband's and wife's postmarriage incomes increase, more savings will occur. This result provides some interesting predictions. Historically the average earned income of women has been about three-fifths that of men. If this gap were to close, couples facing divorce in a Cournot world would save less than they do now. The model also suggests that if two couples have identical levels of postdivorce income, but one is a dual-career couple and the other a couple in which the wife is a full-time homemaker with a low postdivorce income stream, the dual-career family will save less.

When the savings equilibrium condition is not met, there will be no marital savings. In this case $c_1^h=c_1^w=\Sigma y_1^i/2$ during the marriage and $c_2^h=y_2^h$ and $c_2^w=y_2^w$ after the divorce. The partner with the higher postmarriage income will have a higher lifetime utility.

In summary, if a couple face divorce and play a Cournot game in which each acts as a follower, family savings will be smaller when state property division laws give either spouse an overwhelming share of marital property; when, for any given level of θ, postdivorce incomes of the spouses are equal or near equal; and when the divorce-income stream weighted by the respective division shares approaches or equals the value of marital income.

In the Stackelberg leader-follower game, equilibrium consumption values are as follows:

$$c^h = \frac{1}{2}\left[\sum y_1^i + \frac{2y_2^h}{1-\theta} - \frac{y_2^w}{\theta}\right]$$

$$c_2^h = \frac{1}{4}(1-\theta)\left[\sum y_1^i - \frac{2y_2^h}{1-\theta} - \frac{y_2^w}{\theta}\right] + y_2^h$$

$$c_1^w = \frac{1}{4}\left[\sum y_1^i - \frac{2y_2^h}{1-\theta} + \frac{3y_2^w}{\theta}\right]$$

$$c_2^w = \frac{1}{4}(\theta)\left[\sum y_1^i - \frac{2y_2^h}{1-\theta} - \frac{y_2^w}{\theta}\right] + y_2^w.$$

This time the condition for savings to be positive is that the following inequality be satisfied:

$$(3) \qquad 2y_2^h/(1-\theta) + y_2^w/\theta \le \sum y_1^i.$$

The larger the difference between the right- and left-hand sides of the above inequality, the greater the savings. Because the husband has been assumed to be leader in the model, he is able to act on the additional information that he has relative to his wife's reaction function; as a result, his marital consumption will be higher, given any values of θ and incomes. Overall, savings in the Stackelberg world will be lower than that produced by the Cournot game as long as (3) is satisfied, which is the likely case. Savings are less because the husband spends more, *ceteris paribus*, while the wife spends less, but not enough to offset her husband's increased spending. In the Stackelberg game, when savings are zero, the couple will behave as they would in the Cournot world; that is, $c_1^h = c_1^w = \Sigma y_1^i/2$ and $c_2^h = y_2^h$ and $c_2^w = y_2^w$.

Given the importance of θ in determining savings in both the Cournot and Stackelberg models, a comment on its value in the fifty states is instructive. In community property states where each party receives an equal share of marital property at divorce, θ equals 1/2 with certainty. In most other states, θ is based on the division principle of equity. Each party's share is based on their contribution to the marriage; thus $\theta = y_1^w/\Sigma y_1^i$.[4] The Cournot and Stackelberg models suggest that community-property states and equity states with a broad definition of income have division rules which encourage marital savings.

To derive the results for the Nash-bargaining model when $\Sigma y_1^i \ge \Sigma y_2^i$, equation (1) can be reexpressed so that each party is maximizing[5]

$$(4) \qquad U^i \le \left(c_1^i + c_2^i\right)/2.$$

In this case, the sum of the utilities of the husband and wife from equation (4) is expressed as

$$(5) \qquad U^h + U^w \le \left(\sum y_1^i + \sum y_2^i\right)/2.$$

This total must be divided between the two by their selection of marital and, therefore, divorce consumption. If the spouses agree, then the Pareto optimal solution set are all pairs (U^h, U^w) satisfying equation (5) and can be obtained by setting $c_1^h = c_2^h = U^h$ and $c_1^w = c_2^w = U^w$.

If the couple cannot agree on how to divide total utility, one possible threat is that each will try to spend as much as possible during the marriage before the other can do so. Thus this threat is parallel in concept to the no-savings outcome of the two previous models. Given such a threat, the threat point will be

$$(6) \qquad T^i = \left[\frac{1}{2}\sum y_1^i y_2^i\right]^{1/2}.$$

If the threat is carried out, each will have only their own income to consume in the second period.

[4] This simple concept rapidly complicates. Some states consider only money income in y_1^i; some include both money and implicit incomes of household production. Further, many states take into account circumstances that might disadvantage a spouse's postmarriage income-earning capacity (for example, presence of young children, ill-health, limited previous work experience, etc.) in setting the value of θ.

[5] For any positive numbers "a" and "b," the geometric inequality $(ab)^{1/2} \le (a+b)/2$ holds. Note the equality holds when $a = b$.

Given the threat, the Nash-bargaining solution has to satisfy a symmetry condition that any gain from bargaining is to be equally divided, that is, $U^h - T^h = U^w - T^w$. Thus conditions (5) and

$$(7)\quad U^h - U^w = \left[\frac{1}{2}\sum y_1^i y_2^h\right]^{1/2} - \left[\frac{1}{2}\sum y_1^i y_2^w\right]^{1/2}$$

must be met. Solving for consumption levels, we obtain

$$c_1^h = c_2^h = U^h = \left(\sum y_1^i + \sum y_2^i\right)/4 + \left[\frac{1}{8}\sum y_1^i y_2^h\right]^{1/2} - \left[\frac{1}{8}\sum y_1^i y_2^w\right]^{1/2}$$

$$c_1^w = c_2^w = U^w = \left(\sum y_1^i + \sum y_2^i\right)/4 - \left[\frac{1}{8}\sum y_1^i y_2^h\right]^{1/2} + \left[\frac{1}{8}\sum y_1^i y_2^w\right]^{1/2}$$

Marital savings, when the couple play a Nash-bargaining game, depend on the difference between the sum of their first- and second-period income streams. Consumption levels for each depend on the difference in the spouse's levels of postmarriage incomes. When $y_2^h > y_2^w$, the husband will spend more than the wife in both periods. Note that the parameter θ is not an influence in the Nash-bargained solution. The absence of θ suggests that through bargaining, the couple can choose their own decision rule and thereby avoid their home state's rule.

One interesting feature of the Nash-bargaining model is that it offers an explanation of the conditions under which alimony payments might be an attractive arrangement for the couple. In situations where one partner prefers the state's division and the other would be better off carrying out the threat, conflict will arise. The conditions under which one spouse prefers the threat, the other the state's division, will vary depending on the values of θ and their incomes. Alimony appears as a result of bargaining when the spouse with the lower second-period income stream voluntarily agrees to reduce first-period consumption in return for a higher postmarriage consumption than could be reached otherwise. This arrangement will occur only if the high-income spouse agrees to transfer some second-period income (alimony) to the other after the divorce. In short, in the bargained solution, the spouse with the worst future prospect (lower y_2^i) agrees to curtail marriage consumption in exchange for alimony after the divorce. Finally, as expected in a Nash game, the bargained solution provides both spouses with higher lifetime utilities than the threat outcome. Other threats would lead to other bargained solutions and could be explored in further work.

III. Summary and Extensions

This paper has focused on the state's marital property rule, the relative size of each parties' postmarriage incomes, and the noncooperative game the couple plays. The interaction of these has provided some insight into the effect of divorce laws on family savings patterns and the optimal behavior of individuals facing divorce.

A number of extensions of this tightly structured analysis are possible. First, the entire discussion has been limited to a situation with no rate of return to savings, no discount rate between the two periods, and a utility function which forces consumption in the two periods towards equality. Inclusion of time preferences and return to savings would interact with the results presented to provide a wide range of possible marital behaviors. The introduction of these elements into the model might also provide some empirically testable hypotheses. For example, given θ, are there fewer divorces in a state when the interest rate is higher? Has the average length of marriages which end in divorce changed as interest rates and the postmarriage incomes of women have changed?

Second, the discussion has been limited to a study of the optimal behavior of spouses considering divorce. Another natural extension would be to introduce the probability of divorce and allow the couple to play a game with mixed strategies that includes the always married outcome. A third expansion would be to reverse the issue and ask, what is the optimal value of θ to minimize divorce, given the future income streams of groups of couples? I am exploring elsewhere these issues as well as the definition of marital property in the fifty states and the laws' rationale in setting values for θ. In light of the high divorce rate in the United States, further analysis of the effect of divorce on consumption patterns of couples appears fruitful.

THE APPROPRIATE RESPONSE TO TRADE BARRIERS AND "UNFAIR" TRADE PRACTICES IN OTHER COUNTRIES

Responding to Trade-Distorting Policies of Other Countries

By ROBERT E. BALDWIN AND T. SCOTT THOMPSON*

A continuing dissatisfaction on the part of the United States with the rules and dispute settlement procedures of the General Agreement on Tariffs and Trade (GATT) has recently focussed attention on the various steps that individual nations can take to reduce, deter, or offset the trade barriers and "unfair" trade practices of other countries. This paper briefly considers from an economic viewpoint the appropriateness of alternative responses to such trade-distorting measures. The first section examines the social welfare standards that are implicit both in the articles of the GATT and most national trade legislation, and explains the currently permitted responses to foreign trade-distorting practices in terms of these standards. The next section considers the effectiveness of the responses allowed under present GATT rules, particularly those relating to increases in tariffs and subsidies by other countries. The final section proposes certain changes in these response rules that are aimed at improving the prospects for the continuance of an open international trading regime.

I

Economists evaluate trade policies mainly in terms of their effects on allocative efficiency and thus real income at a national or world level. They generally assume that internal redistributive goals are met through nontrade tax and expenditure policies. However, those who frame national trade legislation and the rules of such international organizations as the GATT adopt a more complicated set of standards. While they believe in general that a liberal international trading regime is desirable because of its long-run real income benefits, they are also greatly concerned with the short-run redistributive implication of shifts in trading patterns. In particular, as W. M. Corden (1974, p. 107) has noted, they tend to hold the view that it is undesirable to permit changes in the pattern of foreign trade to reduce substantially and rapidly the real income of any significant group in the economy. Since the consumption impact of most shifts in trading patterns is spread thinly over many individuals, whereas the production effect is concentrated upon a relatively small group, this viewpoint imparts a bias in favor of producers into national and international trading rules.

In the minds of most policymakers, the degree of the undesirability of an income reduction to a particular sector depends not only upon the magnitude and rapidity of the decline, but also upon its cause and the nature of the affected economic sector. If, for example, the income loss is due to the deliberate action of a foreign government rather than to a basic shift in comparative advantage brought about by free market forces, national and international trading rules permit a country to respond under a weaker standard of injury, for example, "material" rather than "serious" injury. A decline in income resulting from increased foreign competition is also regarded as more actionable if the injured industry is an import-competing one rather than an export sector. For example, the GATT permits a country to impose countervailing duties against subsidized imports if it determines that they are causing, or threatening to cause, material injury to a domestic industry. However, when an export industry is injured by foreign subsidization, under GATT rules the adversely

*Department of Economics, University of Wisconsin, Madison, WI 53706.

affected country can only undertake countermeasures if authorized by the GATT Subsidies Committee.

Present international and domestic trading rules are also based upon a particular view of the way in which economies adjust to change. Increases in imports are regarded as tending to cause injurious short-run reductions in income and employment in import-competing domestic sectors. In contrast, export increases caused by foreign duty reduction bring about income and employment gains to export sectors. Consequently, negotiators regard their duty reductions as "concessions" to foreigners that must be balanced by comparable reductions in foreign tariffs for it to be worthwhile to reduce their own trade barriers.

Under this framework of thinking, when a country raises a tariff which it had previously cut as part of a formal trade negotiation, GATT rules and practices specify that this country must either reduce tariffs on other items in order to maintain the existing balance of concessions, or else be willing to accept comparable duty increases on other items by its trading partners. The purpose of permitting foreign countries to raise their duties under these circumstances seems to be to discourage such initial tariff increases, since if the objective had been to offset any injury in the affected foreign export sector, the GATT would have allowed foreigners to subsidize exports from this sector.

The GATT rules relating to permitted responses to domestic subsidies by a country are based on a somewhat different line of thinking. In the situation in which a domestic subsidy increases a country's exports and thereby causes material injury to an import-competing sector abroad, the foreign country is permitted to impose a countervailing duty equivalent to the subsidy. The purpose here appears to be to offset the injury in the particular industry rather than to deter the domestic subsidization. As already noted, in the situation in which the subsidization reduces imports into the country from foreigners, a unilateral response on the part of foreigners is not permitted. The subsidizing country must, however, undertake consultations with another GATT member which believes the subsidy either causes injury to its domestic industry or nullifies or impairs its GATT benefits. If a mutually acceptable solution is not reached, the GATT Subsidies Committee may appoint a panel of experts to assess the situation. The Committee may then authorize "such countermeasures as may be appropriate."

II

How well have the responses sanctioned by the GATT to foreign trade-distorting measures worked in practice? As far as tariffs are concerned, the evaluation must be that GATT procedures and rules have worked quite well. The preferred GATT procedure of reacting to foreign tariffs by offering to undertake a multilateral tariff-reducing negotiation has been utilized on a regular basis since the end of World War II, and has resulted in significant reductions in the level of import duties maintained by GATT members. Furthermore, large countries in general have not tried to exploit their power to improve their terms of trade nor to impose profit-shifting tariffs under imperfectly competitive international market conditions. As the traditional terms-of-trade literature and recent analyses of trade policy under imperfect competition have shown, (for example, James Brander and Barbara Spencer, 1982a), it is usually in the national interest of a country to introduce tariffs if foreign governments do not respond in kind. Moreover, even if other governments do retaliate, a particular country may end up better off than under free trade. However, potential world income is reduced in the process. Just why tariff increases based on this nationalistic motivation have not been more widespread is not entirely clear. In part, it may be due to the acceptance of the view, based on the experience of the 1930's, that all countries eventually lose in a retaliatory tariff war.

There are, however, certain developments that threaten the effectiveness of GATT procedures for achieving a continued liberalization of world trade. One is the breakdown of the so-called "escape clause" or safeguards provisions (Article 19) of the GATT. According to this article (and also U.S. trade law), the level of import protection for a domestic industry can be raised if increased imports cause or threaten to cause serious injury to the industry. In recent years many governments have avoided the obligation of having to compensate other countries with reductions in protection on other product lines (or risk retaliation) by negotiating orderly marketing agreements or voluntary export restraints outside of the GATT framework with the particular countries that are the main source of the increased imports. The agreements generally take the form of quantitative restrictions, and compensation is not provided to the restricting countries. These latter countries are persuaded to accept the arrangements because they usually receive the windfall gains associated with the quotas and are also threatened with more severe restrictions if they do not cooperate. Other exporting countries are generally quite satisfied, since they can often capture part of the market lost by the export-restricting countries. However, the outcome from an economic viewpoint is a greater distortion of world production than if the degree of protection had been increased on a most-favored-nation basis.

A second disturbing development is the failure to maintain the downward trend in protection in certain important sectors. For example, many countries continue to restrict severely imports of agricultural and textile products. Some trade policy leaders in the United States believe that the failure by certain countries to reduce tariffs in key sectors and their greater use of nontariff measures to offset previous tariff concessions have resulted in a situation in which U.S. markets are much more open to other countries than foreign markets are to U.S. goods and services. Consequently, they argue that a new concept of reciprocity should be introduced into U.S. trade law whereby the United States would be able to raise its trade barriers against foreign countries that failed to grant access to their markets to the same extent that the United States does to its markets.

The greater use in recent years of nontariff measures to influence the volume and composition of trade has also reduced the effectiveness of GATT procedures for liberalizing world trade. It has turned out to be much more difficult to undertake successful multilateral negotiations that reduce the use of these measures than to cut tariffs. For example, despite the success in negotiating a new subsidies code during the Tokyo Round, there is still much dissatisfaction in some countries over the way in which the GATT subsidy rules are operating. The U.S. trade officials and industry representatives, in particular, maintain that more extensive subsidization by other governments gives industry in these countries an unfair advantage over American firms. They argue that the U.S. government has ignored this subsidization far too long and should now adopt an aggressive policy of countervailing such subsidies with import duties and equivalent subsidies for U.S. exports to third-country markets. The aim of some who favor this approach is to force other countries to the negotiating table for the purpose of formulating tighter GATT subsidy rules, while the purpose of others is simply to outdo these countries at the subsidies game by using the superior resources and market power of the United States.

Some countries maintain that the United States is improperly adopting this aggressive policy to mask a decline in competitiveness and an unwillingness to adjust to the new realities of comparative advantage. They claim that many of their subsidies are aimed at offsetting market distortions or at easing adjustment problems, and therefore should not be countervailed. However, whatever the merits of the various arguments, it is clear that a trade policy disequilibrium situation exists, in which retaliatory episodes of countervailing duties or subsidies may become much more frequent than in the recent past.

Recent analyses of optimal trade policies under imperfectly competitive market conditions strengthen the view that this type of governmental behavior may become more prevalent. Writers such as Brander and Spencer (1982b) demonstrate, for example, that under some imperfectly competitive circumstances, a country can increase its economic welfare by capturing larger market shares in oligopolistic international markets if its government subsidizes exports or production. It is also in the interest of other governments whose firms are competing in these markets to follow suit and subsidize the exports or production of their firms.

III

The developments described above indicate a need for new international efforts to agree upon appropriate responses to trade-distorting policies of other countries. With regard to safeguards procedures and the compensation issue, it is quite possible that countries will over time become dissatisfied with the use of procedures outside of the GATT framework involving quantitative restrictions and the abandonment of the most-favored-nation (*MFN*) principle as they gain more experience with their ineffectiveness and the political ill will they generate. Unfortunately, it may take a number of years for the drawbacks of discrimination to be fully appreciated by the trading community. Consequently, one meritorious proposal aimed at encouraging countries to follow the *MFN* principle in their current safeguards actions is to eliminate the compensation requirement provided the protection is limited in duration, for example, three to five years, and decreases over the period. Other countries would be able to demand compensation or be permitted to retaliate if the protection extended beyond the period.

Changes in safeguards procedures are also needed to encourage governments to employ tariffs rather than quantitative restrictions for protective purposes. Representatives of injured industries generally prefer import quotas over tariffs on the grounds that their effect in restricting the value of imports is more certain. They often do not appreciate the fact that the tendency for foreign suppliers to shift toward higher unit-value varieties of the protected products creates the same type of uncertainty as the lack of precise knowledge concerning the elasticity of import demand. However, from a national viewpoint the main drawback of quantitative restrictions administered by the exporting countries is the transfer to foreigners of all or part of the revenue equivalent to what would be collected from a tariff.

As a means of overcoming the objection to the use of tariffs to provide temporary protection, a fast-track procedure should be established whereby after an affirmative import injury determination has been made, tariffs could be quickly increased if the value of imports fails to decline to the level deemed necessary by the government to help the industry to adjust in the allotted time period. Utilizing the tariff revenues generated from the temporary protection for adjustment assistance purposes in the injured industry, as suggested by Gary Hufbauer and Howard Rosen (1983), would also help to encourage the use of tariffs in preference to quotas. This proposal additionally has appeal on equity grounds, since domestic consumers who benefit from the injury-causing imports would be required to give up a portion of these benefits to finance the costs of adjusting to the imports.

Ascertaining reciprocity by whether levels of protection between two countries are the same rather than, as has been traditional, by whether changes in the levels of protection between two countries are the same would be a major departure from the procedures followed thus far in achieving the post-World War II liberalization of world trade. Multilateral negotiations would be replaced by bilateral negotiations that could result in a different level of protection against each country for any particular product. Not only would this significantly increase the distortion of world trade, but the outcome of such time-consuming negotiations could in many instances be levels of protection that prove ineffective due to shifts in supply sources and in the production stages at which goods are traded.

Those proposing a new reciprocity concept seem to be mainly concerned with what they regard as the absence of a balance of trade concessions between the United States and Japan. However, this issue can be settled within the present framework of the GATT. Specifically, if U.S. officials believe that Japan or any other country has nullified or impaired the trading benefits accruing to the

United States from prior cuts in U.S. and Japanese tariffs by introducing offsetting nontariff measures, they can bring a GATT Article 23 action against Japan. Under this procedure, failing a satisfactory resolution of the dispute through consultation between the two countries, the members of the organization rule on the matter after receiving the report of a panel of experts. Since other countries would undoubtedly utilize the Article 23 dispute settlement procedure should the United States unilaterally adopt the new reciprocity concept, there seems little reason for the United States not to follow this route itself unless it is prepared at this stage to risk the complete breakdown of the present international trading regime.

Changes in the GATT as well as national rules dealing with the responses of governments to subsidization by other countries are, however, very much needed. There is insufficient distinction between subsidies that tend to raise potential real income in the world and those that achieve some national aim at the cost of worsening the allocation of world resources. There is also inadequate recognition under existing rules of the appropriateness of using subsidies temporarily to facilitate adjustment in industries injured by trade-related causes. Furthermore, greater agreement is needed in distinguishing between subsidies that are sufficiently general to be unlikely to cause material injury to a particular industry and those that are countervailable because they do cause such injury.

As economists have long pointed out, the existence of market failures due to production and consumption externalities calls for the use of production and consumption subsidies (or taxes) as first-best policies to offset these market distortions. For example, as Paul Krugman (1983) points out, there is a case for providing government support for *R&D*-intensive, technologically progressive industries because investment in knowledge in these sectors produces knowledge benefits in other firms and sectors. Instead of reacting to such subsidization by imposing countervailing duties, other countries should consider whether research-oriented subsidies are also appropriate for their own *R&D*-intensive industries. However, all subsidies of this sort should be fully reported and justified in the GATT and be subject to panel determinations as to whether they do in fact operate in the direction of raising real income levels for all countries. Furthermore, in judging the merits of this type of subsidization, the GATT should only accept first-best policy offsets to market failures. Otherwise, not only is the direction of the real income change uncertain under second-best policies, but countries may be encouraged to settle for policies that are inferior in terms of the long-run income goals of the community of trading nations.

Current international tensions over subsidization could also be eased considerably if subsidies to some industries were placed within the safeguards framework of the GATT. There seems little reason for not allowing temporary production subsidies to industries seriously injured by increased imports in the same way that temporary tariff increases are permitted for such industries. Furthermore, export-oriented industries seriously injured because of the penetration of other countries into their third-country export markets should be permitted to receive temporary subsidies without being countervailed. However, as in the case when tariffs are employed, the circumstances for such aid should be fully reported and justified in the GATT and strictly limited in duration.

Some progress has recently been made in deciding upon what types of government subsidies should not be countervailable because of their generality. For example, the Office of International Trade Administration in the U.S. Department of Commerce has proposed not countervailing against subsidies generally available to all industries and regions or directed at a broad productive factor or intermediate input. However, more detailed studies of the allocative effects of such subsidies are needed before clear international rules on these matters can be agreed upon.

The general objective of the international trading community should be to establish a self-enforcing behavior framework in which responses by individual members discourage any single member from pursuing actions that distort the allocation of world resources. In a world of large trading powers and imperfectly competitive product markets, the ability to take credible counteraction is essential to achieve such a goal. This should include actions to offset trade-distorting policies directed at both a country's domestic and export markets. However, if broad agreement on such a framework is to be reached, it must be recognized both that some government policies with significant trade effects can improve the allocation of world resources and that governments are often forced to utilize second-best measures for adjustment purposes. Countervailing actions should not be automatic under these circumstances. Because of the use of trade policies for such purposes, there will remain a need under any self-enforcing framework for a quasi-judicial means for settling trade disputes among countries. Therefore, another much needed change is to improve the GATT dispute settlement process by insulating it from the short-run interests of the disputants and raising the level of competency of the panels involved in the decision-making process.

REFERENCES

Brander, James A. and Spencer, Barbara J., (1982a) "Tariff Protection and Imperfect Competition," in H. Kierzkowski, ed., *Monopolistic Competition in International Trade*, Oxford: Oxford University Press, forthcoming.

______ **and** ______, (1982b) "International R&D Rivalry and Industrial Strategy," *Review of Economic Studies*, forthcoming.

Corden, W. M., *Trade Policy and Economic Welfare*, Oxford: Clarendon Press, 1974.

Hufbauer, Gary C. and Rosen, Howard, "Managing Comparative Advantage," unpublished paper, 1983.

Krugman, Paul, "International Competition and U.S. Economic Growth," Discussion Paper, Urban Institute, Washington, September 1983.

Perspectives on the Jurisprudence of International Trade

By JOHN H. JACKSON*

I tackle a problem which I believe concerns all our disciplines; the problem of the legal processes involved in international trade regulation, and its various costs and benefits. Much of what I say could be applied to international processes, obligations, and institutions such as the GATT or OECD, but for reasons of time and space I will generally confine myself to the domestic U.S. laws and procedures concerning imports.[1]

During the post-World War II period, there have been two parallel but clear trends in the system of United States regulation for imports. The first has been for the overall dramatic reduction in the level of tariffs since 1945, after the negotiation of the GATT, and the seven tariff and trade negotiating rounds under the auspices of GATT.

The second trend has been a gradually accelerating recourse to measures for restraining imports other than normal tariffs, including measures entitled "antidumping duties" and "countervailing duties." This trend has particularly accelerated since 1962, and it is instructive to examine the major trade acts of 1962, 1974, and 1979 (the latter being the Trade Agreements Act of 1979, which implemented the results of the Tokyo Round Multilateral Trade Negotiations). The clear trend manifested in those statutes is towards a greater "legalization" or "judicialization" of the system. The 1974 act greatly reduced administrative discretion in the application of certain regulatory principles, particularly countervailing duties. It did this by imposing time limits, and in some cases embellishing the requirements for public hearings and other procedures to allow citizen access to the process. The 1979 act went even further in this regard, and also took some major steps in expanding the scope for judicial review of administrative actions.

Consequently, as of this writing in 1983, the United States has a remarkably elaborate governmental system for the regulation of imports, including approximately a dozen different formal types of procedures or processes, many of which have explicit statutory procedural requirements calling for public hearings, judicial review, citizen complaint, and much reduced discretion for Executive Branch officials handling these matters. (See my 1977 book.) These include proceedings for escape clause, antidumping, countervailing duty, §337 unfair trade actions, §301 complaints against foreign government actions, etc.

It is said that the U.S. legalistic system of regulating trade is costly, is itself a "non-tariff barrier" to trade, and lends itself to manipulative use by special domestic interests. Some of this may be true, but a systematic appraisal must examine at least three questions. 1) What are the real costs of the system? 2) What are the benefits of the system? 3) What alternatives to the system exist or are feasible, and what are their costs and benefits? I will therefore discuss those three questions, along with some policy and historical matters.

I. The Policy Goals of the U.S. Government Institutions and Procedures for Regulating Imports

Obviously the first policy goal of any system is to arrive at the "right decision" in specific cases. The right decision in cases of import regulation is not always easy to ascertain. Perhaps it is to permit the greatest access for imports coming into the United States, that can be accomplished without causing undue harm to the U.S. economy, without being too unjust to particular segments of the U.S. economy, and without upsetting important international relations of the United States with foreign countries. But specific cases are very complex, and often

*Hessel E. Yntem Professor of Law, University of Michigan Law School, Ann Arbor, MI 49109-1215.

[1] Even then I can only give a brief overview of the subject. A fuller version of this paper will be published in 1984.

pose dilemmas between contradictory goals, such as the goal of permitting maximum imports on the one hand, but preventing unfair hardship on particular small segments of the U.S. economy.

Most of the remaining policies that I will enumerate could be categorized as "procedural."

1) The procedure should maximize the opportunity of government officials to receive all relevant information, arguments, and prospectives. Thus, a procedure which allows all interested parties to present evidence and arguments should enhance this goal.

2) The procedure should prevent corruption and ethical mala fides; that is, to prevent "back room deals" which tend to defeat broader policy objectives of the U.S. government, in favor of special or particular interests.

3) The procedure should enhance the feeling of all parties who will be affected by a decision, that they have had their chance to present information and arguments; that is, that they have had their "day in court." This can be an important policy objective, particularly for democratic societies, so that affected parties can have some confidence in the decision-making process, even when the decision goes against them.

4) The procedure should also give citizens a general feeling that it is fair and tends to maximize the chances for a correct decision. A sense of fairness will include a desire that even weaker interests in a society will be treated fairly; that is, that the ability to get a favorable decision does not depend only on money, political power, or status.

5) The procedure should be reasonably efficient, that is, allow reasonably quick government decisions and minimize the cost both to government and to private parties of arriving at those decisions. It is this policy goal that is perhaps most in question under the American "legalistic" procedures.

6) The procedure should tend to maximize the likelihood that a decision will be made on a broad general national basis (or international basis), not catering particularly to special interests. In other words, the procedure should be designed so that government officials can realistically be assisted in "fending off" special interests which conflict with the broader general good of the nation.

7) The procedure must fit into the overall constitutional system of the society concerned, and be consistent with policy goals supporting that constitutional system. For the United States, an important policy supporting the Constitution is the prevention of power monopolies within the society. Thus a system of checks and balances was designed for the Constitution, which creates a constant tension between various branches of the U.S. government—for example, between the president and the Congress.

8) The system should promote predictability and stability of decisions. This predictability of decisions, whether based on precedent, statutory formulas, or otherwise, enables private parties and their counselors (lawyers, economists, or politicians) to calculate the potential or lack of potential for a favorable decision under each of a variety of different procedures, and to avoid initiating processes with little chance of success.

II. Costs of the System—Quantitative and Nonquantitative

It is interesting to appraise the costs to the U.S. society of the U.S. government "legalistic" system of regulating imports. Here I will omit certain costs which would be incurred no matter what sort of import regulation system a government is likely to have, whether it be a system of great government discretion or a more legalistic system with hearings, statutory criteria, and judicial review (for example, normal customs tariff enforcement which all nations have).

Basically these costs can be divided into two types: those which are quantifiable; and those which are difficult, if not impossible, to quantify.

A. *Quantifiable Costs of the Import Regulation System*

As a rather simplistic exercise, I have tried to evaluate the quantifiable costs in dollars of the U.S. method of regulating imports. A careful evaluation would involve a rather costly survey research study, and I have not had the resources to undertake that, nor do I think it is likely to produce results that are meaningfully better than my "rough and dirty" techniques. Basically the quantifiable costs can be divided into three categories: 1) the budgetary costs of U.S. government agencies concerned; 2) the costs of private attorneys and external consultants who handle the cases; and 3) the in-house costs of the firms who are engaged in the various processes. In the space available here, I cannot detail the methods of arriving at these costs (or their limitations), but the following is a brief summary (these are 1983 estimates based partly on 1981 or 1982 figures. These figures are still tentative, but probably overstate costs): 1) Governmental Costs (line budget items for relevant agencies): total: $200 million. 2) Lawyers and External Consultants (confidential information on fees for various procedures multiplied by number of procedures, excluding exceptional "surges" such as the 1982 steel cases): total: $20 million. 3) Internal Firm Costs (a guess, based on inquiries, that they equal the external fee costs): total: $20 million. The overall total is about $240 million. It seems exceedingly safe to say that the total is less than $300 million per year.

What can we compare this figure with? One obvious comparison is the total value of imports during the year, and for 1983 that is estimated to be $254 billion. The result is 0.00118, or approximately one-tenth of 1 percent. One could conclude that this is reasonably insignificant, if it were considered as sort of a "transaction cost" for a regulatory system which had other benefits. It is perhaps not entirely fair, however, to measure or evaluate the cost of the system by dividing those quantifiable costs by the total value of imports. A better cost-benefit approach would be to look at the welfare benefit to society of the regulatory system. (See below.) It should also be recognized that this aggregate approach does not answer all relevant questions. For example, the distribution of costs can vary enormously, and may in fact be very unfair (imposing, for example, substantial burdens on certain sectors of the economy, and few burdens on other sectors).

B. *Nonquantifiable Costs of the Import Regulatory System*

To focus only on the quantifiable dollar costs of the system, would be a major mistake. Some of the most important costs may in fact be "nonquantifiable." A few of these should be mentioned.

1) *Foreign Policy Rigidity.* A system that depends on statutory criteria and procedures, allows citizen access, and establishes predictability, will inherently diminish the discretion and flexibility of government officials. Indeed, that is exactly what it is designed to do. However, certain types of foreign policy activities may be inhibited by such a system. Secret and delicate negotiations are much more difficult, and quick decisions are sometimes impossible.

2) *Manipulation/Harassment.* The legalistic type system that exists in the United States also lends itself to some abuse by special interests who manipulate the system for their own advantage in ways not necessarily contemplated by the Congress when it enacted the relevant statute. For example, a complainant may be tempted to bring a procedure knowing that it will have considerable opportunity to create mischief and difficulty for U.S. foreign policy through the procedure, but using the procedure as a device for negotiating with the government towards some solution desired by the complainant that is not contemplated within the statutory or regulatory procedure set up by Congress.

3) *Wrong Law Rigidity.* One of the results of the American "legalistic" system of regulating imports is that criteria tend to be embodied in statutes enacted by Congress, and then become very hard to change. On some occasions, the statutory formula proves later to be inappropriate. In these cases it has proved very difficult to get the Congress to change the law, because a variety of special interests tend to be able to block such change.

4) *Special Interest Influence on the Formulation of the Statutory Criteria.* The

processes by which the Congress writes the statutory criteria and formulates the law sometimes lends itself to manipulation by special economic interests in the United States who can foresee the results of certain statutory wordings on their potential cases in the future.

5) *Big Cases Mishandled.* One of the allegations often made is that the legalistic system of the United States for regulating imports may operate with reasonable satisfaction as to the little cases, but when it comes to very big cases which have a broad influence in major sectors of the economy (such as autos, textiles, agriculture, steel), the system breaks down, and in fact what can be observed is a return by one subterfuge or another to a "nonrule system" of extensive executive discretion and "back-room bargains."

6) *The Dilemma of a Legalistic System.* As one can begin to surmise from analyzing these various costs, both quantifiable and nonquantifiable, that to a certain extent there is a dilemma involved in designing any institutional system for regulating imports. This dilemma is that the more one maximizes the goals of a legalistic system (predictability, transparency, corruption minimization, minimization of political back-room deals), the more one will sacrifice other desirable goals such as flexibility and, in international relations, the ability of government officials to make determinations in the broad national interest as opposed to catering to specific special interests.

III. The Benefits of the System

The benefits of the "legalistic system" may be considerable, but they are harder to appraise. I will discuss them under two categories.

A. *Procedural Benefits of the System*

The legalistic system responds well to many of the goals and objectives set out in Section I above, and for reasons of time and space constraint I will not go into much detail here. Clearly, the more extensive and detailed are the statutory criteria, the public proceedings, opportunity for judicial review, and the rest, the more likely that the system will be predictable, corruption proof, minimize back-room political deals, etc. An exception to this might be the "big cases."

B. *Substantive Benefits*

One of the critical questions is whether this legalistic system, given its costs, in fact provides a substantial measure of benefits (benefits which exceed the costs) to the general welfare of the United States. If the legalistic system in fact allows a higher degree of access for imports into the U.S. economy, and if such trade liberalization provides a benefit to the U.S. economy, then one can appraise this benefit. Alan Deardorff and Robert Stern (1984) have used their very large international trade model to compute some of the welfare benefits of liberal trade. For example, they conclude that a 50 percent reduction in pre-Tokyo Round tariff levels across the board would result in an additional welfare benefit to the U.S. economy of approximately $1 billion. This is also a symmetrical result—that is, a 50 percent increase of tariffs causes a comparable decline in welfare.

Many people believe that the U.S. legalistic system—cumbersome, rigid, and costly as it is—in fact provides for an economy more open to imports than virtually any other major industrial economy in the world. If we agree, we can count this as a benefit. How could we measure that benefit? That is obviously very difficult. We are measuring it against an unknown—namely, what would be the degree of import restraint into the U.S. economy if the U.S. system were not so legalistic and were more "discretion prone"? The current import restraints are fairly modest in comparison with other societies, so one might well imagine that the tariff equivalent of import restraints unfettered by a legalistic system might be about 50 percent above current levels. Thus the Deardorff-Stern welfare benefit amount of $1 billion might be one "ballpark" measure of the more quantifiable of the economic benefits of the trade regulatory system. This compares favorably to the quantifiable costs of about $300 million.

One must not forget, however, that there are also a number of nonquantifiable benefits to the system: greater confidence of the citizenry in the operation of the government in this subject matter; the advantages generally for business planning of a higher degree of stability of governmental actions; reduction of corruption; etc.

IV. Some Concluding Remarks and Perspectives

What I have tried to do in this paper is briefly approach the question of whether the U.S. "legalistic and procedural system" of regulating imports has, despite its considerable costs, advantages which outweigh those costs. It is very difficult to be too precise, at least as to the quantifiable aspects, but even indulging in overassumption of quantifiable costs, and an underestimate of quantifiable benefits, one can see that the benefits appear to be very substantial compared to the costs. However, the most important part of the subject may indeed be the nonquantifiable parts, and on those one is likely to receive many different opinions. In short, the matter seems to be very much "judgmental."

One thing is clear, however. Those who would criticize the existing system must bear the responsibility of weighing that system against viable alternative systems. It is not enough simply to describe in great detail all the "horribles" or detriments of the existing system. It is necessary to weigh in the balance the advantages of the system, *and* to do that in comparison with viable alternative systems that might be put in place. Is there any viable alternative system that is likely to be as satisfactory or even more satisfactory than the existing system? (Modest modifications of the existing system are also possible!) What are the possible alternative systems? The principal one that seems to come to mind is one that would involve a considerably higher degree of government official discretion. We can witness such a system in other major industrial countries with considerable imports. Such observation does not lead one to be confident about those alternatives to the U.S. system. The dangers of corruption are high; and often the weaker segments of the domestic economy (usually including consumers) are the ones who must pay for the resulting decisions that are made for the benefit of the more powerful interests. The legalistic system assists well-intentioned governmental officials to "fend off" certain types of particularistic pressures (but of course, no system will fend off all such pressures). Sometimes governmental officials, past, present or future, express considerable impatience with the U.S. legalistic system, and yearn for a "simpler" system. Often they are simply expressing a bias that can frequently be perceived in government officials, that a system should leave to those government officials as much discretion and "elbow room" as possible to make the necessary decisions, because those officials inherently feel that they will make the best decisions possible. Others of us may not have such a high degree of confidence in government officialdom.

REFERENCES

Deardorff, Alan and Stern, Robert, "The Structure and Sample Results of the Michigan Computational Model of World Production and Trade," presented at the Symposium on General Equilibrium Trade Policy Modelling, Columbia University, April 5–6, 1984.

Jackson, John H., *Legal Problems of International Economic*, Relations, Cases, Materials and Text on the National and International Regulation of Transnational Economic Relations, St. Paul: West Publishing, 1977.

Unfair Trade Practices: The Case for a Differential Response

By JUDITH L. GOLDSTEIN AND STEPHEN D. KRASNER*

Trade-distorting practices by other states did not seriously affect U.S. commercial interests until the mid-1960's. However, growing overall and sectoral trade deficits, as well as rapid changes in trading patterns, have made such practices a more salient political issue. The modal and preferred American policy response has been to rely on U.S.-supported liberal institutions, notably the General Agreement on Tariffs and Trade (GATT), to provide a framework for multilateral negotiations designed to eliminate such practices. Although GATT negotiations have sharply reduced tariffs, certain nontariff barriers (*NTB*s) continue to have a deleterious impact on American corporations and more general national economic interests.

To further long-term objectives related to both American economic prosperity and an open global system, two strategies should be adopted. First, the United States should pursue a Tit for Tat strategy with states that violate explicit GATT rules. Conciliation and unimplemented threats, the mainstay of existing policy, will not work. Rather, the United States should retaliate against such violations. Second, for nonconventional *NTB*s, that is, practices not understood to be nor easily capable of becoming incorporated into the GATT, the United States should abandon its policy of attempting to conclude new agreements that would broaden the scope of international rules. The United States has had and will continue to have limited success with such a policy. As an alternative, the United States should develop a more efficacious set of domestic industrial policies, even recognizing that given American institutions and values, this will not be an easy task.

In short, we make the following argument. American interests lie with the continuation of a liberal trading regime. The United States can best insure international cooperation by responding in kind to foreign practices. If states are willing to cooperate, so should the United States; if states defect from accepted GATT rules and norms, the United States should retaliate. If the practice is halted, American sanctions should be stopped. Alternatively, when U.S. economic interests are threatened by nonconventional *NTB*s or export-promoting practices, the United States should accept such behavior as inherent in the trading system and consider adopting similar practices of its own. The principled rejection of industrial policy must be replaced by a pragmatic assessment of options given American institutional and material resources.

I

In conception, the postwar trade regime was to regulate all trade distortions. Liberal rules, norms, and procedures were to be adopted by states and patrolled by international organizations. Problems with this image, however, were evident from the start. The International Trade Organization was never created. In GATT's formative period, 1947–58, two deviations from liberal design appeared: First, a series of national protectionist practices were sanctioned through a "grandfathering" provision. These included the U.S. escape clause provision and the 1921 Anti-Dumping Act. Second, although regime norms were accepted as the prototype, strict adherence to rules and procedures was not immediately expected.

Between 1958 and the close of the Kennedy Round, deviations from liberal norms were tacitly sanctioned in several critical areas. GATT coordinated and monitored the reduction of tariff barriers, but some *NTB*s such as the EEC's Common Agricultural Policy (CAP) were considered to be beyond the regime's legitimate control. Although seen

*Assistant Professor and Professor, respectively, Department of Political Science, Stanford University, Stanford, CA 94305.

as violating the spirit of GATT, such practices were not challenged. For U.S. policymakers, bolstering growth in noncommunist economies took procedence over strict adherence to liberal norms.

By the early 1970's, however, America's concern about discrimination rose. The overall balance of payments situation deteriorated and sectoral problems multiplied. While many of these changes reflected macroeconomic factors, including the effort to maintain ambitious social and military activities in the late 1960's, others were related directly or indirectly to state activities in other countries. New issues, such as the impact of foreign subsidies on the American position in third markets, became more sensitive.

The GATT has been ineffective in dealing with a wide range of problems, especially those not directly related to explicit tariff barriers. Its dispute settlement mechanism is weak and infrequently utilized (John Jackson, 1979). Many forms of state intervention which impact on international market performance are not directly dealt with in the General Agreement. The GATT secretariat does not even have a mandate to maintain a comprehensive inventory of nontariff barriers. (Under Article X, GATT members are obligated to publish their trade regulations but not in any systematic comparable manner, see Andrzej Olechowski and Gary Sampson, 1980.) Thus, as trade tensions increased in the 1970's the fundamental weaknesses of this organization, originally viewed only as a temporary bridge to the International Trade Organization, became apparent. The U.S. response to these problems was to press for the inclusion of nontariff barriers into the GATT framework in the Tokyo Round of multilateral trade negotiations (*MTN*).

II

With regard to the relevance of present international rules there are three kinds of practices that influence international transactions: tariff barriers, conventional nontariff barriers, and nonconventional nontariff barriers. Tariff barriers have been the primary concern of the postwar trading regime. Several rounds of multilateral trade negotiations have resulted in dramatic reductions.

Conventional nontariff barriers are those that are accepted as being subject to the rules of the international regime for trade. Some, such as quotas, are explicitly dealt with in GATT. Others, such as some voluntary export restraint agreements, have come under the panoply of GATT during the postwar period. The defining characteristic of these conventional *NTB*s is that policymakers recognize them as being on the agenda of the international regime for trade. Some consensus exists as to what constitutes fair and unfair practices.

Nonconventional *NTB*s, however, present a more difficult problem. Nonconventional *NTB*s are practices that have not been accepted as part of the agenda for the trading regime. Basic standards of fairness are subject to debate. The issues involve not how a particular activity should be classified, but the classification scheme itself. The right of one state to challenge the legitimacy of an explicit export subsidy or a quota is not questioned, but the right of a state to demand changes in the banking or distribution systems of a trading partner is much more suspect.

The various *NTB* codes negotiated during the Tokyo Round represent the most significant effort thus far to expand the scope of the GATT to make nonconventional *NTB*s into conventional *NTB*s. While it is too early to reach a definitive conclusion about the success of these agreements, initial experiences are not hopeful. For example, the long and caustic debate between the United States and Japan over purchases of telecommunications equipment associated with the new code on government procurement has had very limited results. Not only has Japan's Nippon Telephone and Telegraph had considerable difficulty opening itself to American producers, the U.S. government was involved in American Telephone and Telegraph's rejection of a low cost bid from a Japanese company for fibre-optic equipment, a decision justified in terms of national security (John Lane, 1983). Moreover, the new codes themselves, whose provisions are limited to the signatories, represent a departure from the fully multilateral principles of the General Agreement. Further, the grievance settlement procedures of the codes have done little to revive the authority of GATT. For example, the first two cases filed under the new subsidies code, initiated by Australia and Brazil, claimed that the EECs sugar export subsidy program was in violation of GATT Article XVI. Although the EEC subsidy was clearly at odds with GATT standards, no sanction was imposed (Jeffrey Estabrook, 1982). And, in practice, had GATT ruled against the EEC, moral suasion would have been the only instrument at its disposal. Thus, while norm-based arguments have led to some changes in the EEC subsidy program, these modifications have not satisfied complainants.

Given the weakness of dispute settlement mechanisms, the United States should adopt a policy of appropriate unilateral trade retaliation as defined by GATT rules rather than multilateral diplomacy in cases involving conventional unfair trade practices. The international system for trade can be regarded as a prisoner's dilemma. The best outcome for an individual player is for that player to cheat (by, for instance, imposing an optimal tariff) while the other player cooperates. However, if both players cheat they will be worse off than if they both had cooperated. The ironic logic of single play prisoner's dilemma, which drives both parties to defeat, is well known. However, cooperation is much more likely under conditions of iterative prisoners' dilemma, but only if both parties are quick to punish and quick to forgive. Experimental findings suggest that the winning strategy for iterative prisoners' dilemma is Tit for Tat (Robert Axelrod, 1981). Tit for Tat is a strategy in which the player cooperates on the first move and then does whatever the other player did on the preceding move.

The United States is in an ideal position to play Tit for Tat because of its large domestic market, provided that both parties are clear about the values in the matrix and the classification of behavior as cooperation (fair) and defection (unfair). With regard to conventional *NTB*s, which are covered by existing international agreements, these conditions are likely to be met. Retaliation by the United States is likely to alter the behavior of trading partners, if those partners understand American demands and regard them as legitimate. However, if the United States fails to defect itself, it can expect continued defection from its trading partners.

Use of a Tit for Tat strategy may be exemplified by a number of recent American actions. In the case of Chinese textiles, retaliation after Chinese refusal to sign new accords led to a return to the bargaining table. Although the short-term Chinese response was further retaliation, a similar American response on the second move convinced the Chinese to cooperate. Similarly, the American response to the Canadian sale of subsidized subway cars to the United States led to a rescinding of the controversial financing program. In two other cases, future actions favorable to the United States may be predicted. Retaliation for EEC export subsidies led to the 1983 flour, butter, and cheese sales to Egypt. By using its own subsidization program, America supplanted a traditional French market and created incentives for negotiations. In a like manner, the U.S. refusal to compensate the Common Market for American import tariffs on speciality steel is likely to lead to European quotas.

Tit for Tat is not aimed at starting a trade war. It is a program that should elicit cooperation and "freer" trade. It is a strategy that could be implemented given existing American institutions, although this would require that American policymakers change their basic attitudes toward retaliation. For the last thirty years, the executive-centered decision-making structure for trade has viewed legal rules and procedures within the United States that protect American producers from foreign unfair practices as a threat to free trade. Laws, such as antidumping and countervailing duty which were enacted before America's liberal period, have been seen as particularly troubling compared with safeguards passed after 1934, because these statutes do not give the executive any discretionary authority (Goldstein, 1983). As opposed to escape clause procedures where the president can negate a ruling by the International Trade Commission (ITC), unfair trade practices carry mandatory duties. In the past, the executive response to this situation has ranged from attempts at persuading industries to retract petitions, to *ad hoc* devices such as "trigger-prices," to the extensive use of the waiver provisions for countervailing duty cases provided under the 1974 Trade Agreement Act. In general, the executive's strategy has been to undercut American actions for fear of stepping onto the "slippery slope" which leads to spiralling retaliation and the end of an open trading system.

This strategy has been misconceived. The defense of liberalism by the most liberal solution to trade distortions has not worked in practice for it does not work in theory. The United States should meet protectionism in kind. We *cannot* defend liberalism unilaterally; without pressure, our trading partners *will not* act in accordance with GATT norms. Thus, the removal of both dumping and countervailing duty responsibility from the Treasury to the Commerce Department is a progressive step toward further international liberalization. Commerce is less reluctant to use existing legal provisions against unfair foreign trade practices.

III

The third group of barriers to trade cannot be treated as the other two. This group is composed of practices that have not come under even minimal regime control. They have been so excluded because they are difficult to quantify and compare; they are often viewed as part of the individual state's sovereign rights. Development of Anglo-America state-society relations made both the United States and Great Britain the logical progenitors of a liberal international regime. Both polities have historically deemphasized the need for a positive state role in economic affairs. The late industrializers and nations which needed to restructure their economies after World War II had no proclivity to accept the model of a *lasissez-faire* state. State intervention into a range of economic activities was accepted.

Most American central decision makers have had difficulty comprehending cross-national variations in state-society relations. Other states were expected to follow the American model: liberal with regard to both international exchange and domestic social groups. This is not a tenable position for guiding policy. Neither international negotiations nor unilateral retaliation will easily convince other states to change basic institutional characteristics of their political economies. It is futile and possibly counterproductive to attempt to negotiate the removal of nonconventional *NTB*s. Even with the expenditure of substantial resources, the United States can expect at best only symbolic changes. The Common Agricultural Policy (CAP) of the EEC and the Japanese industrial policy illustrate these propositions.

The EEC's agricultural policy covers 90 percent of Europe's farm production. It guarantees producers a uniform international price, usually above the world market price. To make products internationally competitive, the EEC must use export subsidies. In the early 1960's, the United States had little problem with the CAP; European integration was considered to be in America's interest and the CAP was an unfortunate but necessary part. By the 1970's, however, U.S. acquiescense to the CAP faded. Not only did the subsidies program undercut American agricultural exports to Europe, but it also weakened the U.S. position in third markets.

The U.S. strategy has been to show that EEC policies violate GATT norms. Although unquestionably the case, negotiation and enactment of a subsidy code during the MTN has had little impact on EEC policy. In affect, the United States is attempting to make the EEC subsidy program a replica of its own, that is, one based on domestic but not international state intervention. This task is problematic. Negotiations on subsidies are difficult to conclude especially with respect to adjudication procedures. Once institutionalized they invite complainants to use GATT's procedures. However, since GATT is incapable of handling such cases, the codes can be counterproductive to their liberal intent. Weak institutional mechanisms further undermine GATT's credibility. And, of central importance, applying pressure to the EEC to change the CAP may threaten general political goals by exacerbating tension between the United States and Europe.

Japanese industrial policy offers an even clearer example of practices that were not envisioned by American policymakers in either the drafting of the GATT or of domestic legislation. Industrial policy refers to state sanctions targeted to specific sectors or industries. Industrial policy can take four different forms: protection, adjustment, relief, and enhancement (Daniel Okimoto, 1983). While protection, adjustment, and relief either fall under existing international rules or are generally consistent with liberal norms, enhancement or nurturing presents overwhelming problems for international regulation or unilateral American initiatives.

The relationship between the state and the private sector in Japan is radically different from the tacit Smithian model upon which American rules have been predicated. The presumption in Japan "is that the state is there to do whatever is appropriate and necessary to promote industrial growth and prosperity" (Okimoto, p. 25). Nurturing has involved specific government support such as low interest funds for targeted industries, direct subsidies, special amortization provisions for capital investments, and exclusion of critical capital equipment from import duties. Although it would be difficult, some of these practices might be directly challenged by the United States. However, there is a wider array of actions that fall under the concept of administrative guidance that reflect even more vividly the unique characteristics of Japan's political economy. The ability of Japan's central economic bureaucracies, the Ministry of International Trade and Industry (MITI), and the Ministry of Finance to exercise administrative guidance comes not so much from the direct allocation of resources, but rather from "respect for the bureaucracy, the ministries' claim that they speak for the national interest, and various informal pressures that the ministries can bring to bear" (Chalmers Johnson, 1982, p. 266). The MITI has acted as a clearing house for information involving the development of national research strategies for critical industries. It has organized Diet caucuses to support specific high technology sectors in exchange for relatively modest political contributions to the Liberal Democratic Party. Major public financial institutions are heavily influenced by MITI preferences. The *keiretsu*, or conglomerates composed of several industrial firms, a trading company, and a bank, which dominate Japan's industrial structure, were constituted by MITI in the 1950's to concentrate capital for key development projects. (see Johnson, Okimoto, and Michael Borrus, 1983).

Neither international arrangements in the GATT or elsewhere, nor American trade policy, can provide an adequate response to Japanese initiatives. Japan is not going to abandon an industrial strategy that has provided such spectacular results in the postwar period. Rather, American pressures could provoke retaliatory or defensive pressures that would threaten market-oriented international behavior. First, Japan would regard such pressures as infringements on existing sovereign prerogatives. One need only contemplate the reaction if U.S. allies demanded that support of defense industries be abandoned because it has provided American firms with commercial advantages. Second, it would be extremely complex to quantitatively specify an appropriate American response. In reflecting on the decision of the Reagan Administration to reject the Houdaille relief petition, the most thoroughly documented case ever presented by an American corporation, Lionel Olmer, the Under Secretary of Commerce, stated that the petition "did make a fairly persuasive case that the Japanese machine tool industry got off to a very successful start on the basis of government assistance of some order of magnitude. The difficulty was to try to allocate that amount of government assistance to a specific product" (JEI *Report*, November 11, 1983, p. 4). In sum, Japanese industrial policy, even more so than the CAP, is not amenable to regulation through international agreement or unilateral international pressure.

The United States should accept deviation from a Keynesian, much less a Smithian, model of state society relations as an inherent characteristic of the international system. Various national policies will cause trade to depart from the pattern that would have developed under purely liberal conditions. Such nonconventional *NTB*s or trade promoting measures are unlikely to change. American interest will not be realized by either expanding the scope of GATT and using Tit for Tat or dispute settlement procedures on these *NTB*s. Rather, the United States, too, may need to rely on "less-than-liberal" solutions to American trade problems. In the coming decades, industrial policy should become one of the states' policy tools. Adjustment to foreign competition, often reflecting the activity of government bureaucracies or publicly owned firms, will require a more active role for the American state. In sum, the United States should learn from the manner in which other nations have undercut American economic supremacy.

IV

In the coming decade, American producers will be increasingly challenged at home and abroad by foreign products. In response, the U.S. should pursue two strategies. When American goods are being supplanted due to foreign use of tariff and nontariff barriers in violation of accepted international norms, the United States should retaliate in kind. A Tit for Tat strategy is optimal. It is the best guarantor of other nations abiding by the "rules of the game." However, a Tit for Tat strategy will not succeed when the trade distortion is not clearly in violation of accepted rules and norms. America's policy in the past in such situations has been to attempt to widen the scope of the liberal regime. However, the United States does not have the power to bring about agreements that would make nonconventional *NTB*s into conventional *NTB*s. Moreover such an effort could be counterproductive. Rather the United States should accept certain deviations from the Keynesian/Smithian model, and when necessary use nonliberal means to bolster the competitive position of United States industries in the international market.

REFERENCES

Axelrod, Robert, "The Emergence of Cooperation Among Egoists," *American Political Science Review*, June 1981, *75*, 306–18.

Borrus, Michael, *Responses to the Japanese Challenge in High Technology: Innovation, Maturity, and U.S.-Japanese Competition in Microelectronics*, Berkeley: Berkeley Roundtable on the International Economy, 1983.

Estabrook, Jeffrey S., "European Community Resistance to the Enforcement of GATT Panel Decisions on Sugar Export Subsidies," *Cornell International Law Journal*, Summer 1982, *15*, 397–427.

Goldstein, Judith L., "A Re-examination of U.S. Trade Policy: An Inquiry into the Causes of Protectionism," unpublished doctoral dissertation, University of California-Los Angeles, 1983.

Jackson, John H., "Governmental Disputes in International Trade Relations: A Proposal in the Context of GATT," *Journal of World Trade Law*, January/February 1979, *13*, 1, 1–21.

Johnson, Chalmers, *MITI and the Japanese Miracle*, Stanford: Stanford University Press, 1982.

Lane, John J., "Phone Fibers, Fujitsu, and the FCC: A National Light at the End of the Northeast Corridor," *Law and Policy in International Business*, 1983, *15*, 653–87.

Okimoto, Daniel I., *Pioneer and Pursuer: The Role of the State in the Evolution of the Japanese and American Semiconductor Industries*, Stanford: Northeast Asia-U.S. Forum on International Policy, 1983.

Olechowski, Andrezej and Sampson, Gary, "Current Trade Restrictions in the EEC, the United States and Japan," *Journal of World Trade Law*, May/June 1980, *14*, 220–31.

Japanese Economic Institute, *Report No. 43A*, November 11, 1983.

Race and Punishment: Directions for Economic Research

By SAMUEL L. MYERS, JR.*

The scholarly debate over the nature and cause of the significant racial disparities in prison incarceration rates in the United States has taken on renewed intensity in recent years. Two sorts of activities have spurred the debate. On one hand, researchers such as Alfred Blumstein (1982), Jan Chaiken and Marcia Chaiken (1982), and Joan Petersilia (1983) have begun to use powerful analytic and conceptual tools to scrutinize the hypothesis that racism or racial discrimination exists in the criminal justice system, or that it is the cause of the racial disproportionality of our prisons. On the other hand, minority scholars and public opinion leaders have begun a very visible and vocal attack on the results of the conventional social science community. (See, for example, National Minority Advisory Council on Criminal Justice, 1980.) These activities have stimulated much discussion among public policymakers and legislators. Ranking black members of the U.S. Congress, for example, have gone on record by questioning social science research findings that purport to show that racial discrimination in certain aspects of the criminal justice system does not exist—or, at least, that its alleged existence is not a cause of the greater representation of blacks in the prisons or the criminal population.

Economists have not been leaders or even active participants in this debate. This is surprising for several reasons. Many of the conventional tools of econometrics can be called upon to resolve some of the statistical issues in dispute; post-Beckerian models are likely to yield more than negligible benefits in sorting out the theoretical effects of punishment on criminal activities; and radical labor market paradigms may prove useful in examining the historical evolution of prisons and punishment in America. A brief overview of a number of different areas of research on race and punishment will illustrate the inherent potential as well as the unrealized promise of economic approaches.

I. Existence and Efficiency Issues

Economists have begun to argue that punishment via incarceration may not be socially efficient. The thrust of the argument is that punishment does not always reduce crime (see my 1983a article). While efforts to deter would-be criminals may be effective, attempts to rehabilitate convicted criminals by making punishment more severe may be futile. For example, in my 1980 article I demonstrate in a two-period rational choice model that criminal-human-capital accumulation, or labor market discrimination, against ex-offenders may result from long periods of incarceration. Rather than reduce the relative return to crime, punishment might increase it. This, in turn, may lead to increased participation in illegitimate activities.

The economic argument can be extended to examine differential treatment of offenders. For example, the certainty and severity of punishment have long been known to differ between blacks and whites. Although the evidence is at times conflicting, it does point to the likelihood that blacks experience harsher punishment than whites.

If punishment cannot be expected to rehabilitate, why then would it be dispensed in relatively more abundant quantities to blacks? Is the answer, perhaps, that black and white offenders differ in backgrounds or personal characteristics? Such a difference, of

*Graduate School of Public and International Affairs, University of Pittsburgh, Pittsburgh, PA 15260. The able research assistance of William Sabol is gratefully acknowledged.

course, could statistically explain any apparent inequalities in punishment. Or, is the answer simply that blacks are more criminal than whites and thus must be punished more soundly?

The issue here is whether racial disparities in punishment (if they exist) are necessary in light of the perceived higher criminality among blacks. Or equivalently, this question is posed: if blacks and whites were punished identically, would crime among black criminals increase? The answer clearly depends on how punishment affects criminality. It also depends on how well one controls for other mitigating factors associated with race, such as employment disadvantage, that may determine criminality.

The standard techniques of "residual discrimination" analysis appear to be quite appropriate for disentangling the effects of race and disadvantage in criminal justice system outcomes. This approach was adopted in my 1981 study.

The results were not entirely surprising. It was found in the analysis of federal offenders that the effects of punishment on black and white ex-offenders differ. Blacks and whites do in fact lower their postprison participation in crime—measured by rearrest rates—after longer prison sentences, but their participation in crime tends to increase when punishment is more certain, measured by the ratio of prison commitments to convictions. Indeed, the crime-increasing effect of punishment certainty is almost twice as large for blacks as it is for whites. However, treating blacks and whites equally in punishment, so that their race-neutral rates of time served and commitment to prison are equated, does not cause an increase in black recidivism. This counterintuitive result is in spite of the fact that blacks serve an average of five more months in federal prisons than they would in a racially neutral sentencing scenario.

The findings also suggest that the incidence of preprison employment disadvantage among black and white federal offenders is approximately the same. In contrast, the effect on employment of various factors such as marital status, mental health problems, and drug and alcohol use differ substantially between blacks and whites. When these differing effects are equalized between the groups, blacks experience better employment. Our findings point, therefore, to possible labor market discrimination against black offenders. Indeed, poor employment histories and other forms of disadvantage may help explain why offenders have high crime rates in the first place.

II. Historical Underpinnings

Other writers have alluded to a legacy of racism as the cause of inequalities in the criminal justice system. They suggest that slavery and its aftermath are at the root of the continuing injustice of longer sentences served by blacks, of their high probabilities of being sent to prison, and generally of the harsher punishment they face in the criminal justice system. Indeed, Thorsten Sellin (1976) has argued that this state of affairs is intimately linked to labor markets: after the Civil War, a loss of a whole class of workers in southern agriculture mandated that the prison system—already evolving as a labor market mechanism—supply public labor since private involuntary servitude had been eliminated.

Sellin's story goes something like this. In the early years of the nation, penitentiaries were designed to house criminals from the master class. Slaves were punished through beatings or execution. Free black criminals were sold as slaves or deported. There was, however, a significant push to make the penitentiaries occupied by the master-class criminals self-supporting, since the costs of imprisonment represented a heavy burden on taxpayers. Why not make the prison turn a profit? In Kentucky this was tried during the early nineteenth century, and the convict-lease system was born. In this system, a profit was made by hiring out the convicts. Attempting to fight the high prices of northern manufacturers and to train machine operators, other southern states, including Louisiana, invited private firms to set up shops in the prisons. Following the Civil War, however, both prison industries and convict-lease systems faced a major challenge in the South. Would these systems apply to the newly emancipated blacks? Would the master class and the former slaves be forced to work side by side? The answer was simple. Since the economy was shattered and there was a rapid outflow of labor from the agricultural sector—where blacks allegedly held a comparative advantage—prisons could be used effectively as a means of continuing slavery. With a system of penal servitude, private slavery would be replaced with public slavery. In part, the Thirteenth Amendment to the U.S. Constitution explicitly authorized "involuntary servitude" as punishment for illegal activities. Southern legislatures rushed to enact legislation and to revise their penal codes, with an almost unbelievably rapid result: within a decade after the Civil War, prison population in the South shifted from being virtually all white to being disproportionately black. And, so the story goes, this is how prisons have become what they are today in America.

The distinguished scholar W. E. B. DuBois can be regarded as a precursor to Sellin's idea of an historical linkage between labor markets and imprisonment. The account that DuBois provided of the evolution of the disproportionate representation of blacks in prisons is apparently based fully on existing census documents and other public records. Lenwood Davis provides a summary of DuBois' account that is amazingly similar to Sellin's characterization:

> After the Civil War the South passed elaborate and ingenious apprentice and vagrancy laws specifically designed to make free blacks and their children work for their former master.... The result of this was a sudden large increase in the apparent criminal population in many Southern states—an increase so large that the states were either unable or unwilling to house and control them properly.... Laws were enacted that authorized public officials to lease the labor of convicts to the highest bidder.... Between 1890 and 1900 it was estimated that about seventy percent of all prisoners in the South were black. [1975, p. 4, based on W. E. B. DuBois, *Some Notes on Negro Crime, Particularly in Georgia*, Atlanta University Press, 1904]

This is a suggestive argument that can be developed and tested rigorously using available census documents.

III. Dynamics and Economic Cycles

Throughout the present century, social scientists have recognized the apparent dynamic stability of the large racial gap in official arrests and incarcerations. Indeed, Sellin, as one of those early writers in 1928, firmly questioned whether the differing economic and social statuses between the races could account for the disparities in crime. Years later, Marvin Wolfgang (1964) would conclude that if blacks and whites faced comparable economic circumstances then their crime rates would converge. Thomas Pettigrew (1964), in contrast, suggested that even if blacks and whites faced identical social and economic opportunities today, the disabilities produced by long years of discrimination and second-class citizenship could still result in continued heavy black involvement in criminal and violent behavior.

Some researchers have viewed this contrast as an entirely statistical issue. What must be done, it is argued, is to use standard multivariate techniques and to control for economic and social class variables; if race has an independent effect on criminal outcomes—arrests, imprisonment, crime involvement, and so on—then race, black culture, racism, or discrimination "explains" the crime gap. Aside from the obvious problem of differentiating among these possible explanations, a nagging problem remains. Have we truly captured the intergenerational, historical legacy phenomena posited? The answer is *no* when cross-section data are employed. It is also likely to be *no* when short time-series are used.

What, if any, theoretical basis could there be for positing an historical phenomenon that drives the changes in crime outcomes? One such theory, which goes beyond Sellin's descriptive account, was articulated in the 1930's by Georg Rusche and Otto Kirchheimer. They argued that the penal system and the form of punishment responded to the system of production. When labor is scarce, there is a tendency to develop a system of punishment that exploits the labor of convicts. In the early stages of capitalist development, they presumed, labor shortages existed necessitating the use of chain gangs and prison work camps.

Ivan Jankovic (1978) has expanded upon this paradigm to suggest that when labor *surpluses* develop, the severity rather than just the form of the prison punishment changes. Longer prison sentences are the prelude to larger prison populations. The size of the surplus pool of unemployed workers, in turn, can be regulated by prison admissions.

The intellectual roots of Jankovic's "labor surplus" notion obviously are found in Marx's *Capital*. Jankovic asserts that the reserve army of labor or the surplus population can be regulated by imprisonment and the social welfare system. He, like William Darity and I (1983) who examined black welfare dependency, recognizes the need for long time-series data to capture the full historical dynamics of what is indeed a process rather than an isolated event. He finds, like Harvey Brenner in his much-quoted study of unemployment and crime (1976), a strong, positive relationship between imprisonment and unemployment.

Yet in the period prior to the rapid influx of black criminals into the federal prison system in the 1960's, Jankovic discerns in the federal data no noticeable relationship between employment and imprisonment. Moreover, he finds that the prison-employment relation is weak during significant periods of recession, although at the onset of the Great Depression the expected positive association between labor surpluses and imprisonment is found. Inexplicably, Jankovic fails to mention the rather extensive labor economics literature revealing that much of the variance in aggregate unemployment rates in nonrecession years can be explained by movements in the unemployment rates of marginal or secondary labor market workers, among whom blacks are disproportionately represented. During a major recession or depression, primary group workers are added to the unemployed masses. Either because of the race-class distinction involved in the surplus regulating mechanism of the prisons, or because of the sheer resource constraint faced by the prison system during severe downturns, the imprisonment-unemployment relation ought to be weakened.

Understanding the historical interaction of race, incarceration, and punishment ought to be high on the agenda of any serious research effort on contemporary labor markets and crime. Economic analysts, moreover, may have much to contribute in filling this significant gap in knowledge.

IV. Concluding Comments

Early research by labor economists studying the relationship between unemployment and crime has failed to reveal a strong and convincing relationship between joblessness, on one hand, and participation in illicit activities, on the other. I have argued elsewhere (1983b) that this is largely due to a failure to consider the role of race in the criminal justice system. Important advances in our understanding of the differing criminal and employment experiences of nonwhites and whites can be gained if economic researchers look beyond "rational" labor-crime choices and consider how criminal justice works as an institution. We may discover that racial differences in how individuals are treated in that institution are not part of an entirely exogenous process. It may be that such factors as the level of unemployment among disadvantaged workers actively shape the long-run outcomes observed in the criminal justice system.

REFERENCES

Blumstein, Alfred, "On the Racial Disproportionality of United States' Prison Populations," *Journal of Criminal Law and Criminology*, Fall 1982, *73*, 1259–81.

Brenner, M. Harvey, Testimony, "Estimating the Social Costs of National Economic Policy: Implication for Mental and Physical Criminal Aggression," Paper No. 5, Joint Economic Committee, Washington, 1976.

Chaiken, Jan M. and Chaiken, Marcia R., *Varieties of Criminal Behavior*, R-2814-NIJ, Santa Monica: Rand Corporation, August

1982.

Darity, William A. Jr. and Myers, Samuel L. Jr., "Changes in Black Family Structure: Implications for Welfare Dependency," *American Economic Review Proceedings*, May 1983, *73*, 59–64.

Davis, Lenwood, "Historical Overview of Crime and Blacks Since 1876," in Lawrence E. Gary and Lee P. Brown, eds., *Crime and its Impact on the Black Community*, Washington: Institute for Urban Affairs and Research, Howard University, 1975.

Jankovic, Ivan, "Social Class and Criminal Sentencing," *Crime and Social Justice*, Fall/Winter 1978, *10*, 9–16.

Myers, Samuel L. Jr., "The Rehabilitation Effect of Punishment," *Economic Inquiry*, July 1980, *18*, 353–66.

______, "Racism and the Criminal Justice System," Discussion Paper No. 657-81, Institute for Research on Poverty, University of Wisconsin-Madison, 1981.

______, (1983a) "Estimating the Economic Model of Crime: Employment versus Punishment Effects," *Quarterly Journal of Economics*, February 1983, *48*, 157–66.

______, (1983b) "Racial Differences in Post-Prison Employment," *Social Science Quarterly*, September 1983, *64*, 655–69.

Petersilia, Joan, *Racial Disparities in the Criminal Justice System*, R-2947-NIJ, Santa Monica: Rand Corporation, June 1983.

Pettigrew, Thomas F., *A Profile of the Negro American*, Princeton: Van Nostrand, 1964.

Rusche, Georg and Kirchheimer, Otto, *Punishment and Social Structure*, New York: Columbia University Press, 1939.

Sellin, Thorsten J., "The Negro Criminal, a Statistical Note," *The Annals of the American Academy of Political and Social Science*, November 1928, *140*, 52–64.

______, *Slavery and the Penal System*, New York: Elsevier, 1976.

Wolfgang, Marvin E., *Crime and Race: Conceptions and Misconceptions*, New York: Institute of Human Relations Press, 1964.

National Minority Advisory Council on Criminal Justice, *The Inequality of Justice: A Report on Crime and the Administration of Justice in the Minority Community*, J-LEAA-009079, September 1980.

Black Women, Economic Disadvantage, and Incentives to Crime

By LLAD PHILLIPS AND HAROLD L. VOTEY, JR.*

In recent decades, women have increasingly engaged in serious felony crime, particularly property crimes such as burglary and larceny. Arrests for women for these offenses have been growing at rates three times that for men. In contrast, female arrests for personal crimes have been growing at rates comparable to those for men.

In a previous paper (1982), we have shown that much of the increase in participation in crime by women can be explained by changes in marital status, increased participation in the labor force, and the lack of economic opportunity as reflected by unemployment levels. In our article (1972), we found a strong link between labor market opportunity and crime for youth when race is taken into account. And for mostly male crime in California in our article (1975), we found that economic opportunities for youth provided an explanation for crime while race did not.

Samuel Myers (1983) showed that lack of employment experience has a differential effect on recidivism for black and white males. But women continue to bear unique responsibilities within the home that place their range of choices in a different context from men or male youths, and it is this context that must be examined to explain their changing criminality.

We are struck by the heightened emphasis of the effects of changing marital status and economic opportunities on black women. Many of the points we make in this regard have been made before and are summarized in William Darity and Myers (1983). As we intend to show, circumstances that provide substantial incentives for general female involvement in crime in many cases reduce black women and their households to the choice of poverty or crime. And, welfare measures intended to alleviate poverty may be providing unintended incentives to commit crime for economic gain.

Our approach is to consider a model of labor market behavior in which individuals respond to constraints typical of those situations in which we find women. We focus on women who have chosen to participate in legal work but are constrained, perhaps by the 40-hour week. Some will be overemployed and seeking part-time work, others underemployed and desiring additional work. For both kinds of constrained workers, crime may become a more attractive option. Then we examine women's involvement with the labor market, as affected by marital status and presence of children, focusing on reasons for the relative disadvantage of black women. As an example of how an extensive list of resulting hypotheses may be validated, we provide an exploratory test of our hypotheses linking part-time work and moonlighting to crime by examining the relationship between the supply of hours (legal work) and the decision to participate in grand larceny.

I. A Model of Labor Market Behavior

Picture the choices faced by a typical entrant into the job market in terms of standard neoclassical microeconomics. One chooses between work and leisure, establishing the number of hours one will work for alternative wages offered in the market. Introduce a constraint on the length of the working day and the individual will either work or not, depending on whether the market wage is above the reservation wage.

In the case of women in a traditional household, we can introduce the concept of committed leisure, not really leisure at all, but the time a woman must commit to the responsibilities of managing a home. This

*University of California, Santa Barbara, CA 93106.

reduces her potential for wage labor and modifies the schedule of hours which she will work at a market wage above her reservation wage. Should the woman have income from a husband for use in maintaining herself and the household, her reservation wage will be accordingly higher.

One point that becomes clear from this analysis is that the desired length of the working day may differ substantially from that which is available. Should the woman, with her household obligations, find all available jobs requiring more time than she prefers to work, we would class her as overemployed if she takes a job. Should the jobs she can obtain provide inadequate hours to supply the income she needs and for which she is willing to put in time, she will be underemployed. Women who are heads of households with very small children and no support from husbands or other sources of nonwage income are the most disadvantaged, since they have the dual role of provider and homemaker.

Our society has attempted to alleviate problems of impoverished women by measures such as Aid for Dependent Children (AFDC). Women who are without sufficient means of support are eligible. Unfortunately such support is generally inadequate, leaving women in need of additional income. Furthermore, under the work incentive program (WIN), they also are likely to lose some welfare support if they work. However, if they take full-time jobs, they may be overemployed because they have high levels of committed leisure.

In contrast to wives and female heads of households, never-married single women are similar to single men in that their committed leisure is low. They may wish to work longer hours than possible with a normal working day. These women may be classified as underemployed. If no overtime work is available, their choices to supplement their income are for moonlighting, the underground economy, or crime.

In view of the very narrow constraints imposed on women, the potential for illegal work to provide income becomes inviting. Self-employment, of which crime may be one of the most readily available and remunerative forms, provides a way to choose a work period adapted to need. Furthermore, unreported (illegal) income cannot reduce ineligibility for welfare support. Thus, even when seemingly desirable jobs are available, women may face strong incentives for resorting to crime as a source of support. When we add to this situation the possibility that competition for legal work has produced a high level of unemployment, there are many factors that make crime potentially attractive. Let us consider how the changing circumstances of women have made these theoretical possibilities more relevant.

II. The Changing Status of Women

Our earlier paper (1982) showed how women's role in society has been changing in ways that relate to this model of labor market behavior. From the 1950's to the present, marriage rates have been on the decline and divorce rates increasing. Greater numbers of women have been responsible for their own support than in the past. The consequences have been two-fold. Labor force participation rates for women have been on the increase and, with increased pressures on our economic system to provide jobs, unemployment rates have risen. With the rise in the demand for jobs occasioned by greater female participation and, additionally by the 1950's baby boom, youth and women are competing for many of the same jobs. The result has been an unmet need for employment by women in general, leaving crime as a viable alternative.

In spite of declines in fertility as a consequence of falling marriage rates and rising divorce rates, increasingly, we find single women as heads of households with young dependent children. In all of this, black women are in a relatively disadvantaged position. A much greater proportion of them are heads of households and the great majority are below the poverty level. While many receive welfare support, incomes remain inadequate. Their unemployment rates are higher than those for whites, and, because of the discouragement effect, their participation rates are lower.

Changing trends in marital status present quite a contrast between whites and blacks. We find that a higher fraction of black women have never married, or, having married, are separated or divorced. Their population percentage with children under age 18 is higher for every marital status (never married, divorced, separated, and widowed), except for married with husband present than is the case for whites. The consequence is that 40.5 percent of black families are maintained by women compared to 11.6 percent for whites. Median family income is substantially higher for white female-headed families than for blacks ($15,341 vs. $9,753). Almost twice as many black female-headed families are solely supported by their own earnings than are whites (16.6 vs. 8.8 percent) and almost twice as many are supported solely by welfare than are whites (32.8 vs. 19.50 percent). To a much higher degree, white women are supported by other nonwage income along with their own earnings than blacks (71.0 vs. 49.6 percent). A key difference between black and white families maintained by women is, that despite the greater poverty of black families, a smaller percentage of the women work, 57.6, compared to 74.6 percent for whites. This is a marked contrast to the participation of wives, with children and husbands, where 72.7 percent of the blacks work compared to 63.8 percent of the whites. The implications of many of these differences by race are discussed in Darity and Myers in their investigation of the reasons for the existing marital status of black women. Our focus is on the effects that difference in marital status has for participation in illegal work, which may be more attractive than legal work for many black women heading families. This is a possibility we intend to explore in future empirical work.

In this paper, we focus on women participating in legal work, but who may be underemployed or overemployed. Thirty percent of workers holding two or more jobs are women. Moonlighting has been on the increase for white women, but not for blacks. The percentage of employed women holding two or more jobs is 3.7 percent for whites and 2.0 percent for blacks. Moonlighting is more attractive to employed women never married and to those widowed, divorced, or separated than to wives. Almost half of the women who moonlight hold two part-time jobs.

An increasing fraction of women have been seeking part-time (less than 35 hours) work. About 54 percent of part-time workers are women and three-fourths of these are married. The percentage of married white women seeking voluntary part-time work is 11.2, compared to 7.7 percent for blacks. For both wives with children and female heads of families, a larger percentage of white women than black work part-time.

The theory suggests crime as an attractive alternative if the legal labor supply is constrained. Since employed black women are less involved in moonlighting and part-time work than white women, we would expect a weaker association between their supply of legal hours and criminal careers, than for whites. For white women, we would expect a higher participation in property crime for part-timers (less than 35 hours), and for those working many hours, than for those working 40-hour weeks. The knowledge that arrests of women for property offenses has been rising at three times the rate for men over the past 25 years makes it clear that women have been responding to changes in their opportunities as theory would predict. What remains to be determined is whether there is a differentiated behavior between black and white women in committing crimes for income, and whether such behavior can be explained by changing economic opportunities. As a part of a planned broader study, we consider some preliminary summary statistics for grand theft.

III. The Supply of Hours to Legitimate Work and the Decision to Participate in Grand Theft

Individual observations from the youth cohort of the *National Longitudinal Survey* were classified by sex, race, hours worked at a legitimate job in the survey week, and self-report of the number of thefts over $50 during the past year. Hours worked were classified in three categories: 1 (1–34 hours), 2 (35–48 hours), 3 (49 hours or more). The number of thefts were classified in two categories: 0 (0 thefts) and 1 (1 or more thefts).

The two-way classifications by theft and hours worked are listed for white and black females, and white and black males in Table 1.

TABLE 1—THEFT AND HOURS WORKED

Hours Worked	No Theft	Theft[a]	Total
A. White Women Only			
1–34	1153	19	1172
35–48	846	12	858
49+	94	5	99
	2093	36	2129
B. Black Women Only			
1–34	256	9	265
35–48	195	3	198
49+	18	0	18
	469	12	481
C. White Men Only			
1–34	1004	96	1100
35–48	874	62	936
49+	249	24	273
	2127	182	2309
D. Black Men Only			
1–34	270	34	304
35–48	250	20	270
49+	39	4	43
	559	58	617

[a] One or more thefts of at least $50.

Hours supplied to legitimate work and the decision to participate or not in grand theft may be jointly determined choices by the individual. However, it is instructive to look at the supply of hours conditional on the decision to participate in theft, and to look at the decision to participate in theft, given the supply of hours. The data presented here have been limited to individuals who decided to participate in legal activity and hence are conditional on that decision.

A. *The Supply of Hours Given the Decision of Whether to Participate in Grand Theft*

Refer, for example, to Table 1, panel A for white women, of those not committing theft, 55 percent supply 34 hours or less, that is, work part-time in the survey week, 40 percent supply 35–48 hours and 4.5 percent supply 49 or more hours. In contrast, a considerably larger fraction of thieves (5/36 = 13.9 percent) supply 49 or more hours. This pattern holds weakly for white men, but not for black men or women. A few white thieves appear to have worked long weeks.

This observation for whites is consistent with an association between those who would be underemployed at a 40-hour week and participation in illegitimate activity. The latter is among a number of options for supplying additional labor, including overtime and moonlighting.

B. *The Decision to Participate in Grand Theft or Not, Given the Hours Supplied to Legitimate Work*

Refer again to the portion of the table for white women, of those supplying less than 35 hours. The percent who report no grand theft is 98.38, and 1.62 percent report one or more thefts. The distribution is similar for women working 35–48 hours, that is, 98.6 percent report no thefts and 1.4 percent do. However, a considerably higher proportion of women who work 49 hours or more report thefts, that is, 5.05, while 94.98 percent report none. For all categories of hours worked, a slightly higher percentage of black women report grand thefts than white women, that is, 2.49 compared to 1.69 percent. However, there are no black women working 49 hours or more who report grand theft.

For all categories of hours worked, white men (7.88) report a higher percentage of grand theft than do white women (1.69). The pattern for percentage of white men reporting grand theft is U-shaped as hours worked increases, 8.73 percent for category 1 (0–34 hours), 6.62 percent for category 2 (39–48 hours), and 8.79 percent for category 3 (49 and more hours). The same U-shaped pattern is observed for white women and black men. For all categories of hours worked, 9.4 percent of black men report grand theft compared to 7.88 percent for white men.

The right side of the U-shaped pattern reflects the association between those who work long weeks and theft, whereas the left side indicates a possible association between part-time workers and grand theft. Thus both those who would be overemployed and those who would be underemployed at 40 hours per week include a higher proportion who report grand theft. A work week constrained to 40 hours can precipitate a search for more work among the underemployed and for less among those overemployed. Crime is an option open to both.

IV. The Association Between the Supply of Hours in Legitimate Work and Theft

The statistical significance of the association between hours worked and the proportions reporting grand theft was calculated from the contingency tables using the *Chi*-square distribution and a likelihood ratio test. For white women, the null hypothesis of no association can be rejected at the 10 percent level. There is no significant difference between the contingency table distribution for white women and black women. The null hypothesis of no association can be rejected at the 20 percent level for white men and at the 30 percent level for black men.

With the data tabulated to control only for sex and race, the association between legitimate hours supplied and the decision of whether to participate in grand theft is weak. However, the same pattern of association holds for different groups and is consistent with the theoretical expectation that workers dissatisfied with a 40-hour work week might be tempted by crime, either as a solution to underemployment or to overemployment. In fact, the participation rate for grand theft varies in a U-shaped pattern with the hours supplied to legitimate work, with the minimum at the interval around 40 hours.

These results suggest that our planned future study of the many and varied factors of marital status, employment, and illegal behavior within our postulated theoretical framework will reveal much about the causes of crime by race and sex.

REFERENCES

Darity, William A. Jr. and Myers, Samuel L. Jr., "Changes in Black Family Structure: Implications for Welfare Dependency," *American Economic Review Proceedings*, May 1983, *73*, 59–64.

Myers, Samuel L. Jr. "Racial Differences in Post-Prison Employment," *Social Science Quarterly*, September 1983, *64*, 655–69.

Phillips, Llad and Votey, Harold L. Jr., "Crime, Youth and the Labor Market," *Journal of Political Economy*, May/June 1972, *80*, 491–504.

______ **and** ______, "Crime Control in California," *Journal of Legal Studies*, June 1975, *N(2)*, 327–49.

______ **and** ______, "The Changing Criminal Activity of Women: A Labor Force Participation Perspective," presented at the Western Regional Sciences Association Meetings, Santa Barbara, February 1982.

A DECADE OF GREATER VARIABILITY IN EXCHANGE RATES: AN ASSESSMENT

Assessing Greater Variability of Exchange Rates: A Private Sector Perspective

By MARINA V.N. WHITMAN*

If I were to characterize what has happened to our views of exchange rate behavior during the eight years since I last surveyed the issues in this forum in 1975, when the era of managed flexibility was still new, I would say that we know more and enjoy it less. Our models have become far more sophisticated and comprehensive, but the improved understanding has been accompanied by a sense of disillusionment (at least for those of us who were enthusiastic supporters of the new system) that things haven't worked as well as we expected, yet without any clear idea of what alternative arrangements might have worked better. The late 1970's and early 1980's were characterized by frequent sudden changes in the global economic and political environment associated with sharp and often persistent shifts in both nominal and real exchange rates. And, even if the long-run effects of such movements on wages, prices, output and incomes are indeed neutral, their short-to-medium run consequences for the international and intersectoral distribution of resources are not. Thus, a heightened concern pervades the international community as to whether the positive aspects of exchange rate flexibility in adjusting to changes in equilibrium conditions outweigh these negative effects of exchange rate volatility on both the national and global economies.

The experience of the past decade has made it clear that, rather than providing insulation against external disturbances and enhancing the autonomy of domestic macroeconomic policy, the move from pegged rates to managed flexibility has simply rerouted the pathways by which disturbances are transmitted internationally. Under pegged rates, for example, the effects of divergent macroeconomic policies were transmitted indirectly through the balance of payments and changes in stocks of international reserves, complicating the control of the domestic money stock and thus of achieving "internal balance." Under the present system, in contrast, external disturbances or divergent policies directly affect relative prices, costs, competitiveness and balance on the current account. External constraints have resurfaced in the form of a steepened Phillips curve (i.e., a worsened short-run tradeoff between inflation and unemployment) caused by feedbacks from exchange rate changes to domestic price levels and rates of inflation—often referred to in the literature as the phenomenon of vicious and virtuous circles. The political and economic sensitivity of these processes, together with the increasing importance of the foreign sector in most countries, seems to have made domestic economies more rather than less vulnerable to disturbances originating abroad. Certainly decision makers in the public and private sectors have become increasingly sensitive to excessive variability of nominal and real exchange rates.

The causes of exchange rate movements that overshoot fundamental equilibrium—defined in terms of a restoration of purchasing power parity plus adaptations to any change in the equilibrium real terms of trade—are numerous. Unanticipated shocks to either the real or the financial sector on

*Vice President and Chief Economist, General Motors Corporation.

either the supply or the demand side, unexpected policy announcements or revisions of previous announcements all may produce overshooting of both nominal and real exchange rates when speeds of adjustment differ in different markets for goods and assets (Rudiger Dornbusch, 1976).

Moreover, increasingly integrated and responsive financial markets, combined with sluggish responses in goods markets, have clearly increased the magnitude of the overshooting associated with movement from one equilibrium to another. More fundamentally, the dynamic stability as well as the efficiency of the foreign-exchange market have come under closer scrutiny. But, regardless of how the technical issues are ultimately resolved, the effects of substantial and persistent deviations from equilibrium on the behavior of decision makers are troublesome.

I. Causes of Exchange Rate Change: The New Learning

I can safely leave to others (for example, Jeffrey Shafer and Bonnie Loopesko, 1983) a systematic survey of what "more" we have learned about the behavior of exchange rates over the past decade. I will simply highlight here developments essential to the organization, elucidation and evaluation of the impact of exchange rate variability on decision makers.

The past decade has seen the development of models of exchange rate behavior that integrate the monetarist and the Keynesian traditions. These synthetic models (for example, William Branson, 1983) incorporate as explanatory variables not only such monetarist ones as relative rates of inflation or growth rates of the money supply, but also real variables related to the level and composition of demand or, in Keynesian terms, income and the current-account balance. They also encompass risk-return considerations related to international portfolio diversification, that is, to the relative supplies of and demand for assets denominated in different currencies. These models, which incorporate both goods markets and asset markets, can thus lay truer claim to representing a general equilibrium framework than either the monetary or the elasticities approach taken separately.

Along with this synthesis, we have seen a significant evolution in the definition and complexity of equilibrium. Discussions of equilibrium conditions distinguish carefully between stocks and flows and their interactions. They also distinguish the characteristics of short- and long-run equilibrium, and detail the transition paths by which a system moves from one to the other. As regards exchange rates, the determination of short-run equilibrium values in asset markets is linked, through interest rates, to commodity and factor markets because financial variables affect income, expenditures, employment, and prices. And, because short-term equilibrium values of the exchange rate, given values of these other variables, may yield a nonzero current-account balance, asset flows will occur that in turn affect the long-run stationary-state equilibrium value of the exchange rate, when the current-account balance is zero and all stocks of assets are constant.

We have also learned the importance of the expected future values of exchange rates and other key explanatory variables in any explanation of exchange rate movements. And, along with recognizing this central role of expectations, we have become acutely aware of the importance of unexpected developments in determining changes in exchange rates. Indeed, under the assumptions of rational expectations and exchange market efficiency, *only* the advent of new information ("news") can cause future spot rates to deviate from the market forecast embodied in the current forward rate, which subsumes all existing relevant information.

As one whose responsibilities include exchange rate forecasting, I find this proposition rather difficult to swallow, especially in light of the substantial and increasing resources devoted to such prognostications. At least two observations offer some solace, however. First, although the evidence from intensive theoretical and econometric work on exchange-market efficiency is still incomplete, the work of Richard Baillie et al. (1983)

and others indicates that the forward rate is a biased forecaster of the future spot rate. This result is consistent with the possibility of modeling uncertainty variables and discovering systematic relationships with lagged variables in the information set, thus revealing new information about the past relevant to the future course of exchange rates. That is, additional regressors drawn from the information set will have nonzero coefficients which will reduce the residual variance of the forecast. However, the cost of developing the expertise necessary to identify relevant information to "beat the market" may be a sensible expenditure only for the larger private sector participants in the market. Second, for forecasts more than a year out, another important concern for corporations like my own, the economic fundamentals that underpin even the simplest equilibrium models begin to outperform the random walk in predicting exchange rate movements (Richard Meese and Kenneth Rogoff, 1982).

These conceptual advances have brought with them significant changes in the explanation of the observed volatility of exchange rates. In the Keynesian view, it was "elasticity pessimism," or the apparent insensitivity of trade flows to changes in relative prices, that called the stability of the system into question. By the mid-1970's, more refined estimates yielded higher long-run elasticities and suggested that the Marshall-Lerner condition was met and that the system was indeed stable. The focus of attention then shifted to market imperfections, including insufficiency of stabilizing speculation, as the cause of the volatility that characterized real world exchange rates. But, with the passage of time, an explanation of these imperfections as temporary phenomena characteristic of a transition period became less and less credible.

More recent theoretical developments now suggest that volatility is inherent in the very nature of the system, caused by rational expectations combined with the frequency and importance of unanticipated exogenous shocks. Fundamentally, the question of the long-run dynamic stability of the exchange rate system remains unsettled. Indeed, most models that incorporate rational expectations are characterized by precarious, knife-edged, saddle-point equilibria, which are uniquely stable only because the explosive path is arbitrarily eliminated. One widely used model, in particular—the nonstochastic portfolio-balance model—also requires a unique initial value for the expected change in the exchange rate to arrive at a locally stable path; without such an "anchor" for expectations, the discontinuous jumps that are required to keep the exchange rate on a stable path might not occur and the rate might not approach equilibrium.

II. Effects on Decision Makers: The Policy Environment

Whether or not the relevant markets are efficient or the system is dynamically stable, governments must still properly be concerned with the social costs of extreme exchange rate volatility and overshooting. But the insights of the late 1970's and early 1980's suggest that effective means to reduce these costs may prove elusive. Sterilized exchange-market intervention, for example, is likely to be at best only marginally helpful. Some scholars argue that, used judiciously and in conjunction with appropriate domestic policies, it may aid adjustment via the expectational effects of "buying credibility" among risk-averse investors, whereas others would deny it even that limited role—arguing that only unsterilized intervention can effectively modify the path of the exchange rate, at the cost of sacrificing control over the domestic stock of money.

Increased capital mobility and high interest elasticities of exchange rate expectations have created an uncomfortable link between developments in financial markets and changes in international competitiveness. Volatility in financial markets has produced volatility in real markets. Concomitantly, in the sphere of public policy, divergences in national monetary policies, and in the mix of monetary and fiscal policies, influence real exchange rates and competitive positions, thus altering the pressures for industrial and trade policies. Persistent overvaluation, for example, can lead to increased pressure for subsidies or import protection, while persistent undervaluation can artificially encourage export activities and reduce the impact of import competition on domestic producers. Conversely, changes in industrial and trade policies alter the pressures on financial variables and monetary policy and contribute to exchange rate overshooting.

Ironically, the move from pegged rates to managed floating has reduced pressures for trade restraints motivated by overall balance-of-payments concerns, while increasing such pressures arising from concern with the current account, with imbalances in bilateral trade flows, and with the sectoral problems of industries particularly vulnerable to changes in their international competitiveness. Obviously, excessive variability of exchange rates is not the only source of such pressures—persistently high rates of unemployment in virtually all industrialized countries also play a key role. Nonetheless, theoretical insights regarding exchange rate determination point to the importance of predictability of economic policy in minimizing rate fluctuations and the social costs associated with them. Increased predictability reduces the size and frequency of "news" that produces sudden jumps in exchange rates and sharp deviations from fundamental equilibrium to which the real economy is forced to adjust.

III. Effects on Decision Makers: Multinational Firms

Extreme variability of exchange rates and the risks associated with that variability have profoundly affected the environment in which multinational firms operate and make decisions about pricing, sourcing, scheduling, financing, and the location of production and investment. And the increase in exchange-related risks has also had differential effects, of course, on the present and expected profits, equity, and return to shareholders of particular firms. Firms have responded to this changed environment by devoting even greater resources to seeking out and developing techniques and mechanisms to insulate themselves as much as possible against the risks to their profitability and viability posed by variation in nominal and real exchange rates.

Actually, a multinational firm faces not one but at least three distinct types of risk associated with changes in exchange rates. The first is the short-term transactions risk, affecting commercial transactions and dividend remittances, associated with changes in nominal rates over a time period too short to allow for compensating changes in the prices of goods sold abroad. A second type of risk, which has no immediate cash-flow implications, is the so-called translation risk associated with the effects of changes in nominal exchange rates on balance-sheet valuations. Finally, and most important, there is the longer-term risk associated with changes in competitive relationships between alternative foreign locations that arise from changes in real exchange rates.

Multinational firms have resorted to a wide variety of methods to manage the transaction and translation risks associated with changes in nominal rates more cost effectively. Some have developed computerized "multilateral netting systems" to offset intrafirm cross-border payables and receivables, and then selectively hedge the resulting net exposures in forward markets or through related money market transactions. Moreover, the fact that firms weigh the cost of cover in relation to their outlook for a particular currency and hedge selectively rather than universally implies that they accept intuitively the mounting evidence that forward rates are not only poor but also biased predictors of future spot rates, and that it may be possible to "beat the market" with expertise. At the same time, the attitude of multinational firms toward the foreign-exchange market appears to be defensive rather then speculative. At General Motors, for example, we have developed an early warning system for monitoring inventory levels and foreign-exchange exposures in "high exchange risk" countries.

Several other strategies are available to firms to reduce current operating risks associated with exchange rate variability, including: leading and lagging of payables and receivables, altering the currency of invoicing, centrally coordinating exchange exposure arrangements and, finally, adjusting domestic or export prices of foreign subsidiaries. In practice, it appears that only the last of these is used to any significant extent (John Blin et al., 1981). Longer-term strategies to reduce accounting exposure to changes in nominal rates include various mechanisms, mainly local-currency borrowings, to balance financial assets and liabilities in a particular currency. Such matching efforts are often limited, however, by legal or institutional constraints, as well as by transactions costs and the underlying operating requirements of the business—and such constraints tend to tighten as a currency weakens. In addition, some of the largest firms are beginning to establish trading company subsidiaries to facilitate soft-currency transactions, countertrade requirements and other multicountry barter arrangements that bypass the foreign-exchange market entirely.

Finally, firms must deal with the risk associated with changes in competitive relationships among alternative foreign locations that arise from changes in *real* exchange rates. Such risks are associated with the location of production and cannot be reduced by strategies affecting the sources of financing. For that reason, most major corporations, including my own, try to focus on forecasts of real rather than nominal exchange rates in making decisions regarding the location of foreign production and investment.

The fact that this last and most fundamental type of exchange risk cannot be mitigated by purely financial strategies suggests that multinational firms' defenses must take the form of modifications in operating behavior. Rachel McCulloch (1983) offers several reasons why the risks associated with floating should tend to promote both vertical and horizontal integration, encouraging substitution of direct foreign investment for trade across international boundaries and thus enhancing the relative importance of intraindustry and intrafirm trade. Changes in the policy environment have tended to reinforce these stimuli, inasmuch as the period of floating rates has been accompanied by a reduction in direct controls on international capital flows but also by increased pressure for trade protectionism and by heightened uncertainty regarding the future trade policy environment. Thus, it is hardly surprising that recent empirical work (Richard Abrams, 1980, and David Cushman, 1983a) is beginning to find evidence of a negative impact of exchange rate variability, both nominal and real, on international trade flows. One study (Cushman, 1983b) also found that such variability had stimulated direct investment, lending support to the hypothesis that firms may substitute such investment for exports when confronted by increased exchange rate risk.

IV. Effects on Decision Makers: Banks

The international operations of commercial banks, like those of multinational firms, have been significantly affected by the unexpected volatility of exchange rates during the past decade. Indeed, despite very active involvement in day-to-day trading of both spot and future foreign exchange as well as money market contracts, banks, in general, do not appear to have been very good forecasters of future spot rates. Although some banks have undoubtedly earned profits from their exchange-markets positions, work by Norman Fieleke (1981) suggests that, in the aggregate, U.S. banks may have failed to earn a gross return on their positions.

In fact, however, banks act essentially as brokers in such undertakings, profiting from transactions while generally attempting to close off positions on an up-to-the-minute basis, and especially avoiding overnight exposures. As a result of this caution—stimulated in part, perhaps, by the well-publicized collapses of the Herstatt and Franklin National banks, as the result of exchange-related losses, at the beginning of the floating-rate period in 1974—banks' net foreign-exchange positions have seldom exceeded 1 percent of their total involvement in a major currency. In contrast, similar positions for nonbanks have typically ranged from about 5 percent in German marks and British pounds to as much as 28 percent in Swiss francs (Fieleke, 1981). Although these findings have not been updated for the 1980's, there is no evidence that, in the aggregate, banks' caution in this respect has been relaxed or that they have increased their net activity in foreign-exchange markets.

As regards their longer-term lending, banks have also indicated aversion to carrying exchange risk, either by denominating loans in their home currency or by matching assets and liabilities in a few selected international currencies. But this does not mean that their loan portfolios have been unaffected by currency fluctuations. The sharp fall of the U.S. dollar in 1971–73, and again in 1977–78, encouraged developing countries to borrow heavily in dollar-denominated liabilities. In 1978, for example, the claims of the nine largest U.S. banks on non-OPEC developing countries amounted to $33.4 billion, or 176 percent of the banks' capital. World recession, rising interest charges and the dramatic rise in the real value of the dollar beginning in 1980 have substantially increased the debt-servicing burden and therefore the demand for additional dollar-denominated debt by these borrowers. By June 1982 these claims had risen to $60.3 billion and 222 percent of capital (Fieleke, 1983).

The commercial banks did not take on the incremental risk without compensation, of course; they began demanding substantial interest premiums over LIBOR from "high-risk" borrowers as early as 1980. Yet the shifting of exchange risk to the borrower has not been without cost to the banks. Although large-scale defaults have so far been avoided, the well-publicized difficulties of a number of large sovereign borrowers and the perceived vulnerability of the international financial system because of the complex linkages of interbank transactions that dominate international lending have clearly had a significant depressing effect on the stock prices and price-earnings ratios of large money-center banks.

V. Conclusion

In sum, we are today both wiser and, at least in my own case, sadder about the behavior of flexible exchange rates today than we were a decade ago. But neither the increased theoretical and empirical wisdom nor the enhanced concern about the effects of exchange rate variability on our economic and public policy environment has produced any emergent consensus regarding an acceptable alternative system. Rather, the stress seems to be on ways to make the present arrangements work better.

In particular, greater credibility and predictability of policy on both the monetary and the "real" side, together with increased international policy coordination, at least in the restricted sense that each of the leading industrialized nations take into account the exchange rate effects in the determination of their monetary and fiscal policies, could help to reduce the volatility of real exchange rate movements around long-term trend values. This damping, in turn, would minimize the misallocation of resources, the pressures for artificial restrictions or stimuli, and the potential for political frictions over the international effects of domestic economic policies. And it would certainly be a powerful boon to firms in the private sector, financial and nonfinancial alike, in their unending quest for ways to manage risk and reduce uncertainty.

REFERENCES

Abrams, Richard K., "International Trade Flows Under Flexible Exchange Rates," *Economic Review, Federal Reserve Bank of Kansas City*, March 1980, *65*, 3–10.

Baillie, Richard T., Lippens, Robert E., and McMahon, Patrick C., "Testing Rational Expectations and Efficiency in the Foreign Exchange Market," *Econometrica*, May 1983, *51*, 553–63.

Blin, John M., Greenbaum, Stuart T., and Jacobs, Donald P., *Flexible Exchange Rates and International Business*, Washington: British North American Committee, 1981.

Branson, William H., "Macroeconomic Determinants of Real Exchange Risk," in R. J. Herring, ed., *Managing Foreign Exchange Risk*, Cambridge: Cambridge University Press, 1983.

Cushman, David O., (1983a) "The Effects of Real Exchange Rate Risk on International Trade," *Journal of International Economics*, August 1983, *15*, 45–63.

______, (1983b) "Real Exchange Rate Risk, Expectations and the Level of Direct Investment," Manuscript, Department of Economics and Finance, University of

New Orleans, January 1983.

Dornbusch, Rudiger, "Expectations and Exchange Rate Dynamics," *Journal of Political Economy*, December 1976, *84*, 1161–176.

Fieleke, Norman S., "Foreign-Currency Positioning by U.S. Firms: Some New Evidence," *Review of Economics and Statistics*, February 1981, *63*, 35–42.

_____, "International Lending on Trial," *New England Economic Review*, May/June 1983, 5–13.

McCulloch, Rachel, "Unexpected Real Consequences of Floating Exchange Rates," *Essays in International Finance*, No. 153, Princeton University, August 1983.

Meese, Richard and Rogoff, Kenneth, "The Out of Sample Failure of Empirical Exchange Rate Models: Sampling Error or Misspecification?," International Finance Discussion Papers No. 204, March 1982.

Shafer, Jeffrey R. and Loopesko, Bonnie E., "Floating Exchange Rates after Ten Years," *Brookings Papers on Economic Activity*, 1:1983, 1–70.

Whitman, M. v.N., "The Payments Adjustment Process and the Exchange Rate Regime: What Have We Learned?," *American Economic Review Proceedings*, May 1975, *65*, 133–46.

Stabilization Policies in Open Economies

By JEFFREY R. SHAFER*

Before the shift in 1973 to floating exchange rates between major currencies, the most widely held views of proponents of floating exchange rates on how economies would behave in such a regime, and what effects monetary and fiscal measures would have at home and abroad were based on simple models. This simplicity was achieved by imposing various strong a priori assumptions. The most common and crucial assumptions were continuous equilibrium in all markets, stable money demand, purchasing power parity, and uncovered interest parity (equality between interest rate differentials and expected changes in exchange rates). While these assumptions were not necessarily expected to hold exactly at all times, they were deemed strong tendencies.

Ten years of floating exchange rates have given ample reason to doubt the utility of models based on these assumptions.[1] The clash between theory and evidence leaves analytical economists without a firm foundation from which to draw policy conclusions. The one major counter to this eroding foundation has been the development of models which emphasize the role of asset markets, and hence of expectations, for exchange rate determination and for macroeconomic behavior more generally. But expectations are not well enough understood to provide models with good records of prediction, which can be used as reliable guides to policymaking.

I. Theory and Policy

The disarray among international macroeconomists would be of little concern outside a small circle of technical specialists if the economic performance of the recent past was not so poor and the prospects for the immediate future were not so uncertain. The combination of poor performance of both the models and the real world has produced unusually large disagreements on economic policy issues, even among those who agree on goals for macroeconomic policy.

The economic problems of the past ten years stem largely from major disruptions in the world economy, most notably two upheavals in oil markets, as well as from failures of economic analysis and policy. These disturbances may ultimately be shown to account for many of the pathologies that have been observed. But this conclusion cannot be taken for granted. Moreover, it cannot be used as an excuse for the models' performances.

Models that are consistent with observed behavior entail more elaborate and complex dynamic interactions among key variables, including expectational variables, and embody fewer strong a priori restrictions than the simple models which commanded widespread support at the outset of floating. But as economists turn to more complicated theories, they face two problems. First, the data do not now provide, and are unlikely to provide in the foreseeable future, enough information to distinguish among alternative theories that may have markedly different policy implications. Second, as models become more complex, the gulf will widen between research economists on the one hand, and policymakers and the interested public on the other hand. The lessons of a model will have to be boiled down to rather simple terms to command broad attention.

What I consider in this paper is how the second problem—the gulf between theory and policy—might be bridged. That would seem to entail reducing the analysis to two dimensions and maintaining a comparative static perspective. Focusing on two policy instruments—monetary policy and fiscal pol-

*Vice President, Federal Reserve Bank of New York, 33 Liberty Street, New York, NY 10045. The views expressed here are my own and not the official views of the Federal Reserve Bank of New York or the Federal Reserve System.

[1] My article with Bonnie Loopesko (1983) reviewed this experience and the largely negative lessons for exchange rate models to be drawn from it. Other related papers are referenced in this article.

icy—in a single open economy provides a way of looking at policy goals, tradeoffs among them, and constraints that may limit their sustained use. At least two instruments seem to be required. The view that monetary policy is all that matters for important domestic and international macroeconomic issues has been undercut by experience with large changes in budgetary policies and more general evidence of the importance of aggregate demand disturbances.

Are two instruments sufficient to study macroeconomic issues in open economies? A number of open economy models provide a role for sterilized exchange market intervention, and some indirect empirical evidence supports the view that sterilized exchange market operations, which do not alter money supplies, could alter exchange rates and consequently other macroeconomic relationships. But direct evidence of very large or very lasting effects from such operations has not been found. While further investigation of sterilized intervention may lead to a more precise understanding of just what scope there may be for such operations, it seems pragmatic at this time to assume that it does not represent a powerful macroeconomic policy instrument, at least when confined to the scale of operations that authorities have heretofore been willing to undertake. As a corollary, exchange market operations that are not sterilized can be viewed as no different from domestic monetary operations.

Reduced-form equations for the endogenous variables of the model, with the policy instruments taken as exogenous, provide a good way to summarize a number of the policy implications of a complex model. However, a reduced-form representation in terms only of policy instruments will not, in general, remain fixed. Equations will shift as a result of exogenous disturbances other than policy changes, and, over time, according to the dynamic properties of the model. Moreover, the representation itself will depend on the time frame of the analysis. But from a given starting point, the effects of specified monetary and fiscal policies on output, prices, interest rates, exchange rates, and trade balances at a specified time in the future or averaged over some future time period can be presented in comparative static terms.

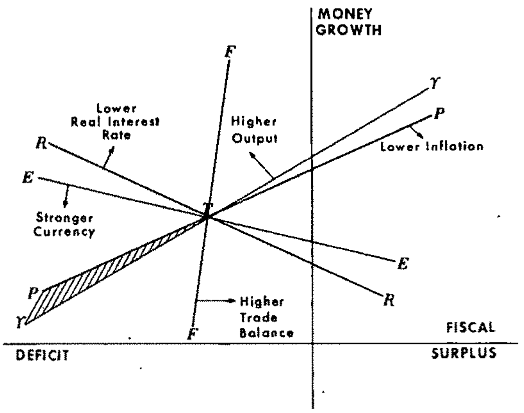


FIGURE 1. SHORT-RUN POLICY CHOICES

The reduced-form equations for a given starting point and time frame can be characterized graphically by iso-value loci for any or all of the endogenous variables. Figure 1 illustrates combinations of monetary and fiscal policies that provide the same outcome for various macroeconomic variables according to fairly broadly held views about key macroeconomic relationships as these views have developed in light of the experience of the past dozen years. These loci should be taken as reflecting relatively short-run effects of policies—say over a period of one or two years. For a specified policy setting T, the positively sloped locus YY represents alternative combinations of tighter monetary policy and looser fiscal policy, and vice versa that would produce the same income. The PP locus also has a positive slope, but it is drawn less steeply, reflecting the assumption that monetary policy, by affecting prices through its influence on exchange rates as well as interest rates, has a greater effect on prices than does fiscal policy for the same effect on output. A downward-sloping locus of unchanged real interest rates RR reflects a tendency for reduced money growth to raise real interest rates and a less strong tendency for a shift towards fiscal surplus to lower real interest rates as it reduces output and prices. The locus EE represents a locus of equal exchange rates. It has been drawn less steeply downward sloping than the RR locus so that the effects of lower real interest rates, which could be expected to be associated with a low (but rising) currency value, and lower domestic prices, which could be expected to be associated with a high currency value, are offsetting as one moves down the locus to the right. The possibility of an upward-sloping *EE* locus will be considered below. A locus of equal foreign trade balances *FF* is drawn to be steeper than the *YY* locus. Moving down along this locus, lower output offsets a stronger domestic currency and lower domestic prices in their effects on the trade deficit. There is no strong a priori reason for this locus to be positively sloped, but if it were negatively sloped it would be steeper than *EE*.

II. Domestic Policy Issues

This framework helps to highlight the range of relationships that might be observed among endogenous variables depending on how policies change or what exogenous variables change and by how much. Policy changes can be represented as movements to a new set of iso-value loci. The members of the new set cannot intersect their corresponding members from the set of loci drawn through the initial point. Exogenous disturbances shift the loci according to the role that the particular disturbance plays in the model. Thus, the framework imposes some discipline on the tendency to relate endogenous variables to one another without regard to underlying exogenous forces. To give one example of this tendency, popular discussions of macroeconomic issues often seem to assume that higher real interest rates imply lower output. This view is a consequence of mistaking a partial equilibrium proposition for a general equilibrium one. The reduced-form approach helps keep the analysis in an appropriate general equilibrium context. Moving away from *T* in most directions would lead to real interest rates and output moving in opposite directions. However, movement into the cone bounded by *RTY* leads to a higher real interest rate and a higher level of output.

Among the important empirical issues that this framework highlights are the implications of the relative slopes of the various loci. To take one implication of the relationships pictured in Figure 1, a larger fiscal deficit combined with a tighter monetary policy offers the prospect of achieving both lower inflation and higher output through currency appreciation. This possibility has attracted attention recently in the wake of large cyclically adjusted fiscal deficits combined with monetary restraint in the United States, while several other major countries have embarked on medium-term programs to reduce fiscal deficits.

Is this course advantageous for the United States, as Figure 1 suggests? One issue raised by this constellation of policies is whether the improvement in macroeconomic performance from such a policy mix entails sectoral burdens that are unfair. This policy mix compresses demand for those goods that are especially sensitive to interest rates and exchange rates more than those that are less so.

A second issue is whether such a policy mix is no more than a beggar-my-neighbor policy in a global economy that is generally plagued with inflation and depressed output. The possibility of reducing inflation and raising output by a change in policy mix assumes no matching change in the policy mix, and a corresponding deterioration of achievable combinations of output and inflation in other countries. For a small country, it is reasonable to believe that policies undertaken abroad would not be significantly influenced by policies chosen at home. Moreover, the impact of domestic policies on foreign economies would be negligible. But a shift in policy by the United States or another large country could be expected to alter adversely the relationship between policies and economic results in other countries. This interdependence raises the question of whether the large countries ought to consider the effects of their policy choices on other countries. I will return to this matter below. Interdependence also has direct implications for the questions at hand since the loci implicitly imbed assumptions about policy reactions abroad. The relationships of Figure 1 assume that policies abroad will be unchanged or respond only moderately to policy changes in the country under study. Policy reactions abroad are likely to reduce and may even eliminate the potential for gain from shifting the policy mix—that is, the *YY* and *PP* loci

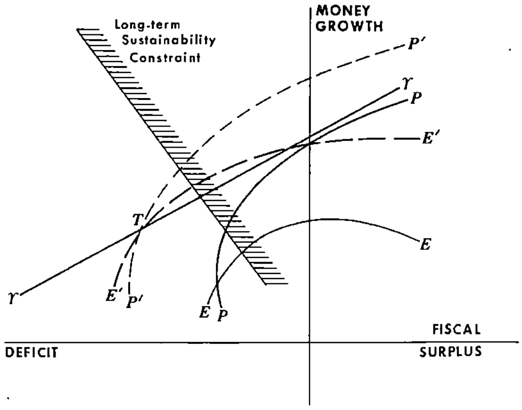


FIGURE 2. CUMULATIVE EFFECTS OF EXTREME POLICY MIX

may more nearly coincide if policies abroad respond to shifting conditions.

A third issue is whether extreme policy mixes are sustainable. From a very long-term perspective, a steady-state growth path would require that nominal government deficits not grow faster than the money stock. Otherwise real interest rates will continue to rise. The capital stock and potential output will grow at less than the achievable steady-state growth rate, and they could decline.

The combination of high fiscal deficits, low money growth, and a strong currency will also generate large current account deficits. The associated rising trend of the countries' net foreign indebtedness could at some point evoke resistance from foreign investors, and the exchange rate would then begin to weaken despite the policy mix.

Figure 2 illustrates how the effects of policy mix of tight monetary policy and large deficits might become less favorable over a longer period of time. The combination of output YY and inflation PP that are attainable at T initially become impossible to achieve over time as the large deficits put pressure on the exchange rate and inflation—that is, the longer-term loci for this variable bend down to the left of a sustainable policy mix. An unchanged policy mix now generates higher inflation P' and a weaker currency E'. The effects on output, if any, are less clear.

III. International Considerations

While the proposed analytical framework highlights the policy choices of only one country, it may nevertheless be useful in clarifying some issues concerning international policy rules of the game.

A major criticism of the fixed exchange rate regime was that its rules precluded countries from pursuing policies that would achieve a desirable level of output and inflation which would otherwise be feasible. Situations such as the one pictured in Figure 3 were seen to present themselves. Here the loci Y^*Y^* and P^*P^* are the optimal levels of output and prices from a domestic point of view. With no exchange rate obligation, the policy choice 0 will produce these results. However, with an obligation to maintain exchange rate E_1, the best policies a country in this situation can pursue are those between points A and B. Compared with the optimum at point 0, income would be lower or inflation would be higher or both. Note that, in this example, the problem is having to keep the currency too low. Given the suggested relationships, which should be considered only transitory ones, the task of defending a high currency value would be less likely to pose a problem since income could be higher and inflation might be lower. The difficulty in such circumstances, if there were one, would be one of sustainability.

One claim of the advocates of floating rates was that such a regime would free up domestic policies to achieve domestic objectives. A stronger claim—that, with floating exchange rates, countries would be insulated from the effects of policies and other disturbances from abroad—has been generally discarded on the basis of the experience with floating to date. But insulation is a stronger condition than is necessary to achieve domestic objectives with unilaterally chosen policies in a floating exchange rate regime. The optimal point 0 in Figure 3 may be moved around by the policy actions of others, as well as by other disturbances and the dynamic evolution of the economy. Nevertheless, so long as such a point exists and exchange rates need not be defended, domestic policies can, in principle, track it.

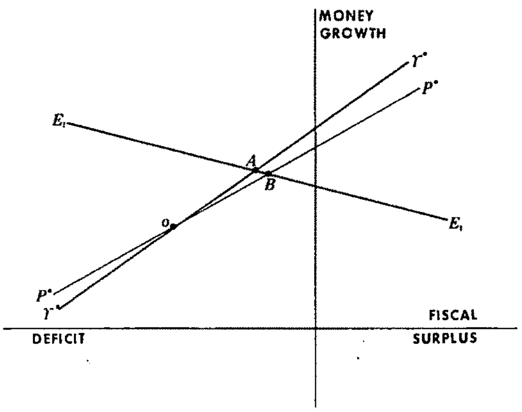


FIGURE 3. FIXED EXCHANGE RATE CONSTRAINT ON POLICIES

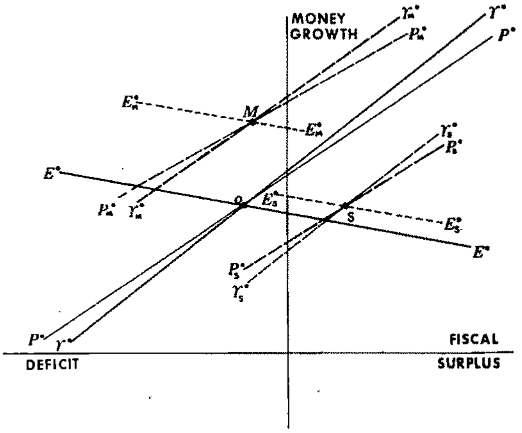


FIGURE 4. EFFECTS OF DISTURBANCES TO MONEY DEMAND AND SPENDING

This conclusion must be drawn with caution, however. The earlier discussion of the sustainability of policies suggests the possibility that because of the policy choices of others, a point like 0 might not exist—P^*P^* and Y^*Y^* may not intersect. A country would then face a choice between excessive inflation and depressed output. This suggests the possibility of a better outcome if other countries cooperated in choosing policies that made the combination of P^* and Y^* attainable. But, in general, such policies would entail some sacrifice of domestic goals in those countries. Moreover, factors other than policies abroad will also influence whether Y^* and P^* can be achieved. These considerations militate against national commitments to rules as a means to cooperative economic policies. However, the possibilities for agreeing on some guidelines for national policies and for pursuing *ad hoc* cooperation will become greater as the implications of policy choices become more predictable.

Even when the theoretical possibility of tracking an optimal set of policies does exist, as a practical matter authorities may find it difficult to do so given disturbances that are unexpected and often not directly observable. This difficulty seems to suggest an alternative rationale for fixing exchange rates or limiting their movement. In the framework sketched above, some disturbances shift the policy mix that is consistent with an unchanged exchange rate by the same amount that the optimal policy is shifted. The exchange rate might therefore provide a useful indicator of the need to alter policies to achieve unchanged domestic objectives in the face of disturbances.

Two such examples are shown in Figure 4. First, a downward shift in the demand for money will shift all three loci vertically to a point like *M*. Such a shift would mean that an unchanged policy will lead to lower output and lower inflation than intended, accompanied by a stronger currency. But an appropriate increase in money would restore the initial prospects for the economy. An upward shift in the propensity to spend on domestic output will shift all three loci horizontally to a point like *S*. This would mean higher output and higher inflation than intended, again accompanied by a stronger currency. In both these examples, the policies that restore the previous output and inflation will lead to the same exchange rate, but the converse is not true—not just any policies that maintain the exchange rate will suffice to restore domestic macroeconomic conditions. Hence, even when the optimal policies are consistent with an unchanged exchange rate, it matters what combination of policies is pursued to maintain the exchange rate. This point receives insufficient attention from those who argue for reducing or eliminating exchange rate fluctuations.

In contrast to these examples, a drop in the risk premium attached to a currency would raise the locus of policies that yield unchanged exchange rates, while the policies that would give unchanged domestic macroeconomic conditions would shift by less. Clearly, the desirability of keying policies to keep the exchange rate unchanged depends crucially on the prevalence of disturbances like the first two relative to the third or on the ability to identify the kind of disturbance that has occurred.

IV. Conclusion

This paper has sought to highlight the kinds of results from theoretical and empirical research that are most relevant to issues of macroeconomic stabilization policy. It has not sought to resolve the crucial questions of how market expectations are formed, how they influence economic behavior, and what are the major exogenous influences affecting the key macroeconomic variables of interdependent economies. Further research is needed to answer these questions and to find theoretical models with durable predictive records. Only with such models can we determine just what the policy tradeoffs are, how they change over time and how they depend on the time horizon of the analysis. If the clouds of uncertainty can be dispersed and new insights are gained, it will be necessary to present policy implications stripped of much technical detail. The approach taken here illustrates one way to do this.

REFERENCE

Shafer, Jeffrey R. and Loopesko, Bonnie E., "Floating Exchange Rates After Ten Years," *Brookings Papers on Economic Activity*, 1:1983, 1–70.

Exchange Rates and Policy Choices: Some Lessons from Interdependence in a Multilateral Perspective

By VAL KOROMZAY, JOHN LLEWELLYN, AND STEPHEN POTTER*

Ten years of floating exchange rates have not resulted in national policy autonomy. Indeed, in today's world, it seems scarcely conceivable that *any* exchange rate regime could enable countries to achieve their domestic objectives independently of what is going on elsewhere in the world; interdependence of national economies simply may not permit independence of national policies. Does this mean that policies directed in each country at getting the domestic situation right are to some extent hostage to the policy choice of others? If so, how do the constraints manifest themselves? Are these constraints made more or less onerous by the way the system works? These questions are addressed by this paper; it would be too much to suggest that they are answered, or, indeed, are answerable in any definitive way.

The early optimism that floating would free countries from balance of payments constraints and thereby enable them to direct policy, particularly monetary policy, to domestic objectives is perhaps understandable. The conditions over the period to the late 1960's had in many respects been particularly favorable to exchange rate stability. It would be natural if this led to the view that the few cases where exchange rate adjustment seemed called for would be better handled if exchange rates were left free to float, so that adjustment could take place relatively early and smoothly. But it would now seem that the typical applied economist or policymaker inherited from the period both a personal data base and a "model" that left him ill-equipped, in a number of ways, for what was to follow.

First, few could have foreseen the extent to which countering inflation would need to become the overriding objective of policy. Second, supply-side shocks became bigger and more numerous. Third, the freedom and volume of financial flows has increased enormously. Fourth, the "system anchor" that had been provided for much of the Bretton Woods period by the *n*th country role and anti-inflationary policies of the United States was lost, and no new anchor put in place.

It could well be that the regime of floating rates that has been in operation over the last ten years has, at least in its broad features, been the only one that could have functioned in the prevailing conditions. If so, it may be that the regime has, at times, had an unwarrantedly bad press. The regime, and arguments put forward for its adoption a decade ago, should be judged not against some hypothetical ideal standard, but rather against what might otherwise have taken place. Ten years on, this judgement is not easy to make: what has not worked well is clearer than what is needed to make things work better.

I. The Transmission of Shocks

Perhaps the main requirement, in assessing the performance of the floating rate regime, is to understand how shocks, both policy and nonpolicy, are transmitted through exchange rates, and the role that such exchange rate movements may play in either damping or amplifying adjustments elsewhere in the system. There has been a noticeable intellectual convergence toward a common analytic framework—the asset-market approach, wherein exchange rates are driven along rather complex paths by the tension between slow-adjusting goods markets and rapidly clearing financial markets. But general, robust propositions about the implications for policy are hard to derive.

*Department of Economics and Statistics, OECD, 2 Rue Andre Pascal, 75775 Paris, CEDEX 16 France. The views expressed are our own, and do not necessarily represent those of the OECD or its member governments.

One proposition that does seem general is that full insulation of the domestic economy through floating exchange rates is an unrealistic special case. It requires either a degree of price flexibility that is not apparent in practice, or a time frame for analysis so long that it discounts the dynamics of adjustment —and associated costs—that macroeconomists have generally taken to be of the essence.

Beyond this, plausible models of exchange rate determination can readily be constructed which have the property that, at least for some shocks, exchange rate responses will tend to amplify overall adjustment costs. Casual observation certainly does suggest that, at one time or another, most countries have experienced situations where the behavior of exchange rates in response to a domestic or foreign shock posed significant difficulties for policy. The capacity, and even the necessary knowledge, to adjust policy to deal with these difficulties was often not in evidence.

In this connection, the more subtle question of what policy "rule" is best does not yield unambiguous answers. Depending on the nature of the shock, and the nature of the economy on which it impinges, different rules can be shown to perform best. In some instances, policy could minimize costs by allowing the exchange rate to absorb the shock; in others this would be achieved by adjusting policy so as to stabilize the rate. Practical application requires, of course, that policymakers can detect the nature of the shock they are facing.

When the analysis is enlarged to consider several countries simultaneously, further ambiguities emerge. Some shocks are absorbed by exchange rate movements in a way that all countries would accept as desirable; but conflict situations can also be modelled where some countries would be better off if exchange rates moved, while other countries would prefer a policy response that prevented, or damped, this movement.

Empirical validation of the various analytic possibilities is another matter. The key point about exchange rates is that they are "system" variables *par excellence*. They are determined in the full set of structural relations defining a linked set of macroeconomic models, and are not readily allocable to a subset of proximate determinants within the overall system. Errors of specification anywhere within such a model structure are likely to show up sharply in the performance of the system in tracking exchange rates in a dynamic simulation.

II. The Properties of the System as a Whole

Consideration of the system as a whole poses questions of both a positive and a normative kind. How does the system as a whole influence the policies, or policy regimes, that individual countries are drawn to adopt? What do these choices add up to at the global level—are they globally optimal or not? And, perhaps most difficult, how would alternative "system choices" affect outcomes for better or worse?

This is an area of considerable importance and controversy at the political level, yet one where the economics profession has not, to date, provided firm guidance. A positive research approach might be to begin by considering why it is that political attitudes about exchange rates differ in quite systematic ways among countries. One point, embedded in the literature on optimal currency areas, is that economic size, and the degree of openness, matter. With a rising weight of international transactions, real and financial, in the economy, the gains from being able to exercise independent policy choices diminish relative to the costs. This literature might lend support to the formation of the EMS, and plausibly give a grounding for the tendency to see the world in dollar-yen-mark terms, but does not really have much to say about whether the relationships among these three rates should be a matter of policy interest.

More recently, game-theoretic approaches to this problem have begun to be explored. One general proposition to emerge—robust though not surprising—is that cooperative solutions to policy setting generally dominate noncooperative ones except under very special circumstances. Much less clear is whether this theoretical dominance is large enough to warrant an internationally coordinated approach to policy, given the inevitable uncertainty and substantial transaction costs that might be involved. More importantly, game-theoretic analysis has opened up in a relatively concrete way the considerations that may lead different players to prefer different strategies, and has identified the possible importance of alternative solution concepts for the distribution of welfare. Differences in domestic economic structure are crucial; the role of real wage flexibility, for instance, has emerged as central not only for analyzing the response of an individual economy to an exchange rate shock, but as a determinant of the kind of rules that policymakers in different countries would prefer.

What more, if anything, is there to be said on the "system" issues? Given the imperfect understanding of this area, any views can only be highly speculative. But two points are worth exploring: the question of international consistency and the "distribution" among countries of policy control.

III. International Consistency

The "*n*th country" problem continues to apply in a floating-rate world. Not all countries can choose their exchange rates, and thereby their terms of trade. But what is by no means clear is the "rule" by which consistency is brought about. One answer, of course, is that no such rule is necessary because countries are not obliged to adopt *explicit* exchange rate or terms-of-trade targets. But events do not automatically work out in a totally satisfactory way for the system as a whole.

One risk, that of a chain reaction of competitive devaluations, was foreseen when the pegged-rate system was disbanded (the memory of the 1930's being quite vivid), and IMF surveillance was adopted as a system safeguard. Exchange rate surveillance is of course easier to apply in a less interdependent world, where the foreign exchange transactions of governments, rather than their overall financial policies, are decisive. In today's world, the threat of competitive devaluation could be more accurately viewed as the risk of globally inflationary policies. These could result if each country were to expand its domestic money supply to counteract currency appreciation and loss of competitiveness stemming from monetary expansion abroad.

The converse is also possible. The endogenous determination of the terms of trade could induce a bias towards global deflation if countries, concerned to reduce inflation, sought to break into the domestic wage-price spiral through the disinflationary effect of currency appreciation and rising terms of trade.

Exchange rate wars, in either direction, remain hypothetical; they have not been observed in the last decade. One answer to this is that implicit commitments (for example, within the OECD) or explicit ones within the IMF framework have been adequate. A second explanation is that countries have been sufficiently uncertain about the relative advantages of a stronger exchange rate (and thereby less inflation), or a weaker one (and thereby a better competitive position) to induce an effective diversity of policy that has, in the end, meant that the compatibility of terms-of-trade objectives has not been an issue. It is also worth noting that in the recent period, with oil and other commodities depressed as a result of weak world demand, the major industrial countries of the OECD have all managed to some extent to achieve favorable terms-of-trade developments.

IV. Distribution of Policy Influence

Different countries perceive differing capacities for independent action. In the early part of the floating period, the demise of the adjustable peg system was often attributed, at least in part, to the relative decline in the economic weight of the United States. It was argued that the period of benign monopolistic guardianship was over, and that unilateral decision making would have to give way to the (admittedly more arduous) process of collective management of the system. Developments through 1979 would indeed have tended to confirm this view—particularly during the period of dollar weakness beginning in late 1977, when talk of a multicurrency reserve system became rife. The Substitution Account exercise might—at least on some interpretations—have been viewed as a way of formally recognizing this "new reality," moving the dollar from center stage.

Events since then suggest that too much was taken for granted (though of course the earlier perceptions may still in the long run prove correct). The United States appears at the moment to have recovered a large degree of autonomy in its economic management. Conversely, in other countries it is the perception of external dependence that seems most in evidence: the inability to achieve domestic objectives because of high interest rates, the strong dollar, the absorption of world savings by U.S. budget deficits, or various other expressions of the constraining influence of U.S. policies. What can economic analysis provide by way of explanation?

Part of the story may be that independence of action stems not so much from economic size as from flexibility in product and labor markets. It is not the whole story, as the recent constraints on action seemingly encountered by the Japanese authorities bear witness, but it is a central point. Currency depreciation is less to be feared if wages do not automatically rise to compensate. Similarly, a change in the terms of trade will have smaller effects on employment if mobility of resources is high. Flexibility thus increases policy autonomy to the extent that the consequences of international repercussions from domestic policy choices are more easily absorbed.

A second explanation might focus on the difficulty that European countries experience in operating as a collective unit. Though Europe as a whole is economically nearly as large as, and little more open than, the United States, each European country is of course considerably smaller, markedly more open and significantly more dependent on external developments. The "psychological" aspect of this perceived dependence, largely true for the individual European economy, may well carry over to situations where (as in the context of EMS), it is not warranted—when it is collective policy action that is at issue.

A third hypothesis is somewhat more technical—and applies more specifically to the present policy environment. It concerns the strength, and duration, of "export crowding out"—that is, the Mundellian proposition that, for given growth of the money stock, a more expansionary fiscal policy will lead to upward pressure on interest rates, an appreciation of the (real) exchange rate and, over time, a widening current-account deficit to accommodate the desired capital inflow resulting from higher domestic interest rates. The role of this mechanism in explaining the current strength of the dollar has received ample attention.

From 1980 to 1983 most countries were concerned to reduce inflation while seeking to minimize the costs in terms of foregone output and employment. The mix of monetary and fiscal policies pursued in the United States may, from that point of view, seem to have been favorable for this country. The high dollar has boosted the terms of trade and helped bring down inflation; meanwhile the associated reduction of net exports has not seemed unduly worrying because total output has nevertheless been rising strongly. Where export crowding out is well established, a mix of tight money and expansionary fiscal policy may thus be seen to have substantial short-run advantages, whatever the longer-term implications for economic structure. But how does this "single country" logic generalize?

If export crowding out were equally characteristic of all economies, each might be tempted by a tight money/easy fiscal policy mix, and associated currency appreciation. But it is evident that the benefits of a stronger exchange rate could not be enjoyed by all. Furthermore, export crowding out may well not be as pronounced in some other countries as it is in the United States. For example, it may be that, because of the relative independence of the Fed, markets are convinced that U.S. deficits will not be monetized, whereas in the case of some European systems of money control, incremental deficits can plausibly be seen as feeding sooner or later into higher money growth. More generally, the export crowding out hypothesis requires that the current account deficit resulting from fiscal expansion will raise risk premia by less than the induced rise in interest rates; the United States may well be a special case in this regard.

If so, there would be no incentive, outside the United States, to use policy asymmetrically. The tradeoff between longer-term costs of an unbalanced mix and short-term benefits from the terms of trade would be unfavorable. And it is noticeable that in practice few countries have sought to emulate the United States policy mix.

To summarize the argument, export crowding out might appear to confer a degree of policy autonomy, to the extent that the asymmetric use of monetary and fiscal policy made it possible to strike a desired balance between domestic demand growth and the terms of trade—to some extent independently of the external environment. In practice, however, it may be that few countries can rely on this mechanism. And even in the United States, export crowding out cannot go on forever. The logic is that eventually the dollar would have to depreciate, not just to restore the trade balance, but to make possible the interest payments that cumulative foreign indebtedness, through the current account, will require. The question, of course, is when and how this correction might take place.

NEW EVIDENCE ON THE DEMAND FOR SAFETY AND THE BEHAVIORAL RESPONSE TO SAFETY REGULATION

Consumer Demand for Automobile Safety

By CLIFFORD WINSTON AND FRED MANNERING*

Given the significant number of fatalities and serious injuries that result from automobile accidents in the United States every year, the issue of automobile safety regulation is of considerable importance. The purpose of this paper is to carry out a cost-benefit analysis of the regulations concerned with the introduction of passive restraints, air bags, and crash-resistant bumpers using improved estimates of the benefits from these safety features. Specifically, our benefit estimates are based on individuals' compensating variations with respect to changes in the probability of severe injury, or expected collision costs that are attributable to the safety regulations. The compensating variations are obtained from a disaggregate vehicle-type choice model of new car purchasing behavior which incorporates safety-related variables in the specification. We present estimation results for the vehicle-type choice model. These results form the basis for our estimates of consumers' compensating variations which are incorporated in our cost-benefit analysis.

I. Compensating Variation Estimates

The vehicle-type choice model used here is a modification of the dynamic model of households' automobile purchasing and utilization behavior we developed earlier (1983). In this analysis, we estimate a logit model of households' utility-maximizing new vehicle purchasing decisions using data from 1980. The choice model is given by the multinomial logit specification,

$$(1) \qquad P_i = \frac{e^{V_i(X_i, S)}}{\sum_j e^{V_j(X_j, S)}},$$

where P_i denotes the probability of selecting a specific make and model (say, Ford Escort) of a new (1980) vehicle i, V_i denotes the mean indirect utility associated with vehicle i, X_i denotes the characteristics of vehicle i (manufacturer-related effects which are captured by vehicle-make dummies, fuel efficiency, capital cost, and safety characteristics), and S denotes household characteristics (income, past vehicle usage, number of vehicles owned). A description of the variables used in the specification is given in Table 1. The primary data source for the household-related variables is the U.S. Department of Energy's Household Transportation Panel which is composed of households included in the National Interim Energy Consumption Survey. The data source for vehicle attributes is a constructed vehicle attribute file. The safety variables are based on safety performance data collected for specific new (1980) vehicles by the Highway Data Loss Institute (1982).[1]

It is important that we comment on the safety variables. In this analysis, we attempt to control for those aspects of automobile safety that specifically reflect damage to the

*Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, and Pennsylvania State University, respectively. We are grateful to the NSF for financial support. A longer version of this paper containing a more detailed discussion of the data, statistical procedures, and calculations performed here is available upon request.

[1] The safety variables used in this analysis are vehicle specific attributes (i.e., exogenous to households) in the sense that procedures are used to correct for vehicle safety performance that is due to differences in the mix of owners who operate specific vehicle makes and models.

TABLE 1—ESTIMATION RESULTS[a]

Variable		Parameter Estimate
Fuel Efficiency (mpg)		.1812 (.0254)
Lagged Utilization		.1101E–03 (.1431E–04)
Vehicle Capital Cost/Income		−5.424 (1.044)
Probability *PI* Claim Exceeds $1,000[b]		−.0564 (.053)
Expected Collision Costs		−.4064E–02 (.1799E–02)
Vehicle Weight		.1140E–02 (.2179E–03)
Vehicle Horsepower for One-Vehicle Households		.0371 (.5985E–02)
Vehicle Horsepower for Two-Vehicle Households		.0263 (.5218E–02)
American Motors Dummy		−.1476 (.6908)
Ford Dummy		.4494 (.2220)
Chrysler Dummy		−.7341 (.2683)
Foreign Car Dummy		.8767 (.3072)
Log Likelihood at Zero	−506.6	
Log Likelihood at Convergence	−363.9	
Number of Observations	220	

[a] Standard errors are shown in parentheses.

[b] Annual probability of a personal injury (*PI*) exceeding $1,000 in *PI* claims conditioned on the occurrence of a personal injury (in percent).

vehicle, and personal injury to vehicle occupants. Note that a variable such as vehicle weight, which we also control for, captures in a general way a vehicle's safety and to some extent its comfort. Thus, our specification includes expected vehicle collision costs and specifies the likelihood of personal injury in terms of the conditional probability of being involved in a severe accident, given that an injury-causing accident has occurred. Although other variables that pertain to personal injury could be constructed, we feel that this variable is more likely to capture households' evaluations of the benefits from vehicle safety improvements that affect occupants' safety as it reflects a highly plausible risk that involves a threat to one's life. We note also that consumers typically have some notion of the value of these variables as they are used to a certain extent in the determination of their insurance rates.

The estimation results are also presented in Table 1. Generally, all of the parameters are of expected sign and of reasonable statistical reliability.[2] It is interesting to note that the dummy variables for all non-GM cars indicate that consumers attach relatively more utility to new vehicles manufactured by foreign companies and Ford. The parameter estimates of most interest are those associated with the safety variables. In particular, it is possible to obtain a quantitative sense about the reasonableness of these parameter estimates by using them to calculate consumers' marginal valuation of the reduction in collision costs and of the conditional probability of severe injury. This calculation yields the results that a representative consumer is willing to pay $17.67 in terms of increased capital costs for an annual $1 reduction in expected collision costs over the life of vehicle ownership, while the consumer is willing to pay $245.30 in terms of increased capital costs for an annual 1 percent reduction (in absolute value) in the conditional probability of severe injury over the life of vehicle ownership. The first estimate is quite reasonable under any plausible assumption regarding the length of vehicle ownership (eight to ten years is typical), and when we consider that it also reflects the value of possibly reducing inconvenience costs involved in a collision (for example, obtaining vehicle repair estimates, loss of vehicle use during repair, and so on). The second estimate is also reasonable as it implies that a consumer would be willing to pay (at mean values of the variables) $4,579.75 (18.67 percent × $245.30) in increased capital costs to eliminate the risk of severe injury from an accident over the life of vehicle ownership.

The parameter estimates of the vehicle-type choice model can be used to provide esti-

[2] We include lagged utilization of similar vehicles (i.e., same manufacturer) to capture the notion of brand loyalty (see our earlier paper).

mates of consumers' compensating variations. As shown by Kenneth Small and Harvey Rosen (1981), the appropriate expression for the compensating variation (CV) in the context of a logit model is given by

$$(2)\quad CV = -(1/\lambda)\left[\ln \sum_{i=1}^{n} \exp(V_i)\right]_{V^0}^{V^f},$$

where λ is the marginal utility of income, V_i is the mean indirect utility associated with vehicle i, n is the total number of new vehicle make and model offerings, and the square brackets indicate the difference in the expression inside when evaluated at the initial and final points.[3] Thus, we can obtain estimates of compensating variations for given changes in expected collision costs or the conditional probability of severe injury that result from specific safety regulations.

II. Cost-Benefit Analysis of Safety Regulations

Using the procedure to estimate consumers' compensating variations presented above, we calculated total benefits from various regulations including the introduction of air cushions and passive restraints, increased usage of manual seat belts (by mandate), and repealing the 5-mile per hour crash-resistant bumper. Net benefits were then obtained by subtracting installation costs from total benefits. The results are presented in Table 2 under a variety of assumptions regarding occupant usage and installation costs (provided by the National Highway Traffic Safety Administration (NHTSA), Ford, and General Motors).

TABLE 2—NET BENEFITS FROM SAFETY REGULATION (IN 1980 DOLLARS)

	Total Net Benefits[a]		
Regulation	NHTSA	GM	Ford
Air Cushion with			
0% Lap Belt Use	6.10	4.33	2.30
20% Lap Belt Use	6.20	4.44	2.41
50% Lap Belt Use[b]	6.76	4.99	2.96
Passive Belts with			
60% Use Rate	5.10	4.80	-
70% Use Rate	5.89	5.59	-
100% Use Rate	9.34	9.04	-
Manual Lap/Shoulder Belts			
50% Use Rate[b]	4.43	4.43	4.43
5 mph Bumper Repeal[c]			
\$100 *CCI*	.31	−.47	-
\$200 *CCI*	−1.40	−2.18	-
\$300 *CCI*	−4.63	−3.56	-

[a] Shown in billions of dollars; based on new car sales of 8.9 million units.

[b] Since this assumption constitutes a significant increase (over current levels) in belt use rate, it might be argued that inconvenience costs should be included in this calculation. One can obtain a rough estimate of these costs by multiplying the time expended annually in buckling and unbuckling belts by the value of time. Under reasonable assumptions, the present value of these costs over the life of the vehicle ownership will not exceed \$1 billion.

[c] Collision cost increases (*CCI*) are increases in average collision losses per accident.

As can be seen, we find substantial net benefits from the installation of air cushions and even larger net benefits (under plausible assumptions) from the installation of passive restraints.[4] Our results differ from Richard Arnould and Henry Grabowski (1981) who find rather low net benefits for air cushions, and from John Graham et al. (1981) who find larger net benefits for air cushions relative to passive restraints. However, the measurement of benefits in these studies was primarily based on estimates of wage differentials in the labor market as opposed to estimates of compensating variations as reflected in the automobile market. It is also interesting to note that we find significant benefits from increased use of manual seat belts (the current usage rate is roughly 11 percent). Finally, we find that repealing the requirement of 5-mile per hour crash-resistant

[3] In our analysis the marginal utility of income, λ, is equal in magnitude (but opposite in sign) to the vehicle capital cost parameter estimate. Since in our specification vehicle capital cost is divided by income, the appropriate estimate of λ is actually the capital cost coefficient (with opposite sign) divided by household income measured in dollars per year. Actual CV estimates were obtained by enumerating through the sample and calculating an average value.

[4] It should be noted that our benefit estimates implicitly control for the vehicle occupancy rate as this is considered by the individual when he purchases a new car. In addition, the personal injury variable is based on observed occupancy rates. Changes in vehicle occupancy rates are not likely to alter the qualitative nature of our findings.

bumpers would generally produce negative net benefits. Repealing this requirement appears to yield positive net benefits only under NHTSA's cost assumptions (as determined under the Reagan Administration).

As has been stated in other automobile cost-benefit studies and should be noted here, the tentative nature of these findings should be stressed. Clearly, these conclusions are sensitive to assumptions with respect to installation costs and the demand analysis that underlies the benefit calculations. Most likely, estimated standard errors for the *CVs* would indicate that more estimation precision is needed to have firm confidence in the benefit estimates. Nevertheless, the approach taken here has proved fruitful in generating plausible estimates of consumers' valuations of important safety devices; as such, it holds considerable promise for successful future application.

In concluding, it is worth noting two issues that have been implicitly raised by this and other cost-benefit automobile regulation studies. First, one is simultaneously struck by the large potential benefits from manual seat belt usage and the current low rate of utilization. A number of explanations have been given to rationalize this phenomenon (see Daniel Orr, 1982, for a summary of these). However, no widely accepted explanation has emerged. Second, it is also noteworthy that although there appear to be considerable net benefits associated with the introduction of air bags and passive restraints (as reflected in consumers' willingness to pay), the automobile companies and to some extent the federal government have not been particularly enthusiastic about these devices.[5] In the case of the automobile companies, it has been argued (see Graham et al.; Peter Passell, 1983) that given the technology involved in the installation of safety devices and the nature of competition in the industry, no company has an incentive to initiate the introduction of these safety features. This suggests that a free market may be incapable of generating the optimal amount of automobile safety protection, thus justifying (even in the absence of externalities) governmental interference. While it is recognized that the likelihood of such interference is often politically motivated, it is nonetheless hoped that if the benefits from safety devices are more firmly established and more widely known, then there will be a greater likelihood that they will be actually realized.

[5] It should be noted that the implementation of government regulations requiring vehicles to be equipped with air cushions or passive belts has been postponed pending a decision by the Department of Transportation as to whether to have the regulations rescinded with supporting arguments that would be acceptable to the Supreme Court or to take steps to have the regulations implemented.

REFERENCES

Arnould, Richard and Grabowski, Henry, "Auto Safety Regulation: An Analysis of Market Failure," *Bell Journal of Economics*, Spring 1981, *12*, 27–48.

Graham, John, Henrion, Max and Morgan, M. Granger, "An Analysis of Federal Policy Toward Automobile Safety Belts and Air Bags," Department of Engineering and Public Policy Working Paper, Carnegie-Mellon University, November 1981.

Mannering, Fred and Winston, Clifford, "Dynamic Models of Household Vehicle Ownership and Utilization: An Empirical Analysis," presented at the 1982 Winter Meeting of the Econometric Society; rev. 1983.

Orr, Daniel, "Incentives and Efficiency in Automobile Safety Regulation," *Quarterly Review of Economics and Business*, Summer 1982, *22*, 43–65.

Passell, Peter, "Airbags for Adam Smith," *New York Times*, May 10, 1983.

Small, Kenneth and Rosen, Harvey, "Applied Welfare Economics with Discrete Choice Models," *Econometrica*, January 1981, *49*, 105–30.

Differences Between Risk Premiums in Union and Nonunion Wages and the Case for Occupational Safety Regulation

By WILLIAM T. DICKENS*

There is an interesting unexplored sideline to the empirical literature on compensating wage differentials (*CD*s) for hazardous work. Every study of differences between union and nonunion compensation for exposure to deadly hazards has found that union members receive much larger *CD*s than nonunion workers.[1] Further, in many of these studies negative *CD*s are found and some are statistically significantly negative.

Some have interpreted these results as indicating the possible existence of substantial market failure. Despite this, there has been almost no discussion of the implications of such a conclusion for occupational safety and health policy. In contrast, several authors, ignoring the union-nonunion differences, have suggested that the empirical evidence on risk premiums supports the argument that markets efficiently allocate occupational risk without government intervention. (See, for example, Robert Smith, 1982, pp. 327, 336.)

The analysis presented below shows that a market failure argument is not needed to explain the finding that union workers receive larger *CD*s than nonunion workers. Efficient contracts may provide workers with either larger or smaller *CD*s than a competitive market would. But, negative *CD*s cannot be reconciled with efficient markets given any reasonable assumptions about workers' preferences. The analysis also considers several potential statistical explanations for these findings. Results are mixed and the conclusion considers the policy implications.

I. The Phenomena

Five studies have examined union-nonunion differences in *CD*s. Richard Thaler and Sherwin Rosen (1976) were the first. In all four of the wage equations they estimated that included a union interaction with risk of fatal injury, the interaction was significantly positive while the risk term was not. Estimated *CD*s were 80 percent to ten times greater for union than nonunion workers. W. Kip Viscusi (1980) estimated fourteen wage equations including sixteen union-risk interaction terms. In all equations, the interaction terms were positive and in twelve cases they were statistically significant. Of the sixteen risk coefficients measuring *CD*s received by nonunion workers, nine were negative and one was significantly negative. Craig Olson (1981), in four specifications, found union *CD*s for deadly hazards were uniformly about six times larger than those received by nonunion workers. Both the nonunion *CD*s and the union interaction term were significantly positive. The *CD*s for frequency of accidents were also larger for union workers. However, estimated *CD*s for union workers for accident severity were negative. A study by Stuart Dorsey (1983) found *CD*s for deadly hazards which were significantly larger for union workers. Nonunion differentials for fatal hazards were significantly negative. Union *CD*s for frequency of nonfatal accidents were much smaller than nonunion *CD*s, while compensation for severity was larger. Finally, Richard Freeman and James Medoff (1981) found *CD*s that were slightly higher for union workers using *CPS* data and smaller using the Em-

*Assistant Professor of Economics, University of California, Berkeley, CA 94720. I thank the Institute of Industrial Relations for generous research support and Chris Martin for research assistance. I have also benefited greatly from the cooperation of James Robinson.

[1] In these studies, compensating differentials are estimated as the coefficients on risk of injury in a wage regression. Theoretical considerations as well as common sense suggest that many omitted variables will be correlated with injury risk. Thus the interpretation of the risk coefficient as a *CD* is dubious. Despite this complication the terms *CD* and risk coefficient will be used as synonyms.

ployer Expenditures for Employee Compensation survey data. All of the *CD*s were positive and none of the differences were statistically significant.

II. Explanations

In general the above results suggest that union *CD*s are larger than those received by nonunion workers. Viscusi has argued that differences are due to the insensitivity of competitive firms to the preferences of inframarginal workers. He suggests that this leads to an underprovision of safety in competitive firms. Olson has noted that larger union *CD*s may reflect unions' informational advantages. Once again the implication would be that occupational safety and health (*OSH*) could be underprovided in the nonunion sector. But, neither of these explanations is necessary. A very simple bargaining model is enough to illustrate this point.

If we assume that a union's objective is to maximize workers' surplus, defined as the difference between the wage and the reservation wage of L identical workers, and that the firm endeavors to maximize profits, the Nash-bargaining solution with bargaining power is the values of L and w which maximize

$$(1)\quad [L(w-r(s))]^{t}[R(L)-wL]^{1-t},$$

where w is the wage, $r(s)$ is the reservation wage which is a function of job safety (s), $R(L)$ is the firm's revenue net of nonlabor costs, and t represents the bargaining power of the union ($0<t<1$).

To determine the *CD*s union workers will receive for more dangerous jobs, we may use the first-order conditions for a maximum with respect to L and w, and the implicit function rule to obtain

$$(2)\qquad \frac{dw}{ds}=r'\left[1+t\left(\frac{R'L-R}{R''L^{2}}-1\right)\right].$$

Equation (2) tells us how wages would differ between two otherwise identical firms which offer marginally different levels of safety. If unions have no bargaining power ($t=0$), the *CD*s received by union workers in the more hazardous firm would be the same as those received by workers in a competitive labor market—r' times the difference in safety. However, if the union has any bargaining power we cannot say whether compensating differentials will be larger or smaller.[2] Thus we do not need a market failure model to explain union workers receiving *CD*s which are larger than those received by nonunion workers. But, the model presented above can not explain negative *CD*s.

All three studies that consider *CD*s for both potentially fatal and nonfatal hazards find some negative risk coefficients. Viscusi finds negative coefficients for both fatal and nonfatal hazards in the nonunion sector. Olson finds negative coefficients for injury severity in the union sector. Dorsey finds a significant negative coefficient for risk of fatal injury in the nonunion sector. Is there any pattern to these findings?

No one finds negative *CD*s for union workers exposed to deadly hazards. Viscusi is the only author who finds negative coefficients for nonfatal hazards in the nonunion sector, and then only in those specifications that do not include a control for deadly hazards. Thus what remains to be explained are the sometimes negative *CD*s for union workers exposed to nonfatal hazards and the negative *CD*s for fatal hazards in the nonunion sector.

Olson suggests an explanation for the first of these puzzles. Because of the protection provided by the union from employer retaliation, union workers may find it easier to get a day off to recover from injuries and may be able to take more time off when they do. Thus the hazard measures used by these three studies—the probability of an accident involving a lost work day and the number of days lost per accident—may not be good measures of the true dangers. If union workers receive some of their compensation for accidents in time off from work, estimated *CD*s could be negative.

[2] For example, if $R=p(Q)Q(L)$, and both p and Q are linear functions, $dw/ds<r'$. If Q is linear but elasticity of demand is constant, $dw/ds>r'$. This indeterminacy is not unique to the Nash solution. For example, a monopoly union model will give similar results.

The finding of negative compensating differentials for fatal hazards for nonunion workers is harder to explain. However, the results are inconsistent. Olson finds positive *CD*s while Viscusi and Dorsey do not. Olson's study includes the fewest controls for other job attributes, but also has the largest sample and the most accurate estimates. It is possible that Viscusi and Dorsey's results are statistical artifacts. It is also possible that adding controls to Olson's study would produce results similar to the other studies. These possibilities are explored below.

III. Additional Analysis

To determine if Olson's results are due to a failure to include controls for other job attributes, several wage equations were estimated. Observations on individual wages and personal characteristics were obtained from the May 1977 *CPS*. Olson also used the *CPS* but used the March and May samples for 1973. The individual data was matched with the BLS data on industry injury and fatality rates for the same year. Wage equations were estimated for union and nonunion workers. The coefficients of probability of a fatal accident and their standard errors are presented in Table 1. Specification 1 replicates Olson's finding of significant positive *CD*s for both union and nonunion workers.

The *CPS* does not contain information on job attributes. To introduce controls for job differences, this study takes the approach of limiting risk comparisons to broadly similar industries. Specification 2 is the same as 1 except for the addition of twenty dummy variables for one- and two-digit industries (risk is measured at the three-digit level). The estimated nonunion *CD* in specification 2 is statistically significant at the 5 percent level. In specification 3, the sample is restricted to workers in manufacturing industries. No industry dummies are included. Once again, the coefficient of probability of a fatal accident in the nonunion sector is negative. Adding ten industry dummy variables again makes the negative coefficient statistically significant at the 5 percent level. Several other specifications were estimated including additional controls for other three-digit industry characteristics. In all cases, measured nonunion *CD*s were negative, and in most cases significantly negative. Thus the inclusion of additional controls reconciles the differences between Olson's findings and those of Dorsey and Viscusi.

TABLE 1—ESTIMATED COMPENSATING WAGE DIFFERENCES FOR EXPOSURE TO DEADLY HAZARDS

Specification	Union Workers	Nonunion Workers
1	6.033	2.116
	(.399)	(.369)
2	3.364	−1.062
	(.651)	(.628)
3	1.536	−1.276
	(.759)	(.833)
4	1.309	−1.564
	(.856)	(.943)

Notes: Fatal accidents per 100 worker years; dependent variable is log hourly wage; standard errors are shown in parentheses. Specification control variables are 1) age, age squared, race dummy, sex dummy, years of education, 3 region dummies, *SMSA* dummy, 7 occupation dummies, probability of an accident involving a lost work day, number of days lost per accident; 2) all controls from 1 plus 20 industry dummies; 3) same as 1 but only workers in manufacturing industries; and 4) same as 3 but add 10 industry dummies.

We may also examine Olson's explanation for negative *CD*s for union workers for nonfatal hazards. If union workers are more likely to get a day off to recover from an accident and get more time off when they are seriously injured, we would expect that an interaction term between the percent of an industry which is unionized and the number of lost workdays per worker would have a negative coefficient. Such a variable was constructed along with a similar one for the probability of a fatal accident. Both variables were added to union and nonunion wage equations with specifications similar to those presented above. The lost-workdays/percent-union interaction was significantly negative in both the union and nonunion equations in all specifications tried. The fatal-danger/percent-union interaction was not. Thus the low *CD*s for exposure to nonfatal hazards may be due to the distortion of the measure of the hazards. Unfortunately, correcting for this distortion would require the estimation of a nonlinear regression. Such an effort is beyond the scope of this paper.

IV. Conclusion

What should we conclude about these findings? First, it is clear that the evidence for the existence of *CD*s is not as strong as past authors have argued. Once we divide the sample between union and nonunion workers, and introduce controls for other job qualities, we begin to find significantly negative *CD*s. The negative union nonfatal *CD*s may be explained by the measurement problems investigated above, but the findings for deadly hazards remain problematic. Arguments that OSH regulation is not needed because observed *CD*s suggest that markets are efficient are not sound. Even if the positive coefficients, sometimes found, reflect risk premiums, it would be inappropriate to argue that employers face the necessary incentives to provide efficient levels of OSH. It is not sufficient that people know that working in a steel mill is more dangerous than working for an insurance company. They must be able to perceive marginal changes in safety within the same firm or between any two firms. If we cannot find significant *CD*s within a broad sector such as manufacturing, it seems imprudent to argue that incentives are adequate.

Second, the volatility of the empirical estimates of *CD*s in these split samples and the theoretical complications introduced by consideration of bargaining and union work rules, argue against another common enterprise—attempts to deduce how people value their lives and job safety from labor market evidence.

Finally, how might we account for the finding of negative *CD*s for exposure to deadly hazards in the nonunion sector? Appeals to imperfect information might explain the lack of *CD*s but cannot explain negative *CD*s. Labor market segmentation might explain negative *CD*s but cannot explain positive nonfatal risk premiums and negative fatal hazard premiums. One is forced to conclude that the available data may simply be inadequate to support investigations of market performance. We will have to look elsewhere for evidence on the efficiency of labor markets and the desirability of safety regulation.

REFERENCES

Dorsey, Stuart, "Employment Hazards and Fringe Benefits: Further Tests for Compensating Differentials," in John D. Worrall, ed., *Safety and the Workforce: Incentives and Disincentives in Workers' Compensation.* Ithaca: New York State School of Industrial Relations, 1983.

Freeman, Richard B. and Medoff, James L., "The Impact of the Percent Organized on Union and Nonunion Wages," *Review of Economics and Statistics*, November 1981, *63*, 561–72.

Olson, Craig A., "An Analysis of Wage Differentials Received by Workers on Dangerous Jobs," *Journal of Human Resources* Spring 1981, *16*, 167–85.

Smith, Robert S., "Protecting Workers' Safety and Health," in Robert W. Poole, Jr., ed., *Instead of Regulation*, Lexington: Lexington Books, 1982, 311–38.

Thaler, Richard and Rosen, Sherwin, "The Value of Saving a Life: Evidence from the Labor Market," in Nester E. Terlecky, ed., *Household Production and Consumption*, New York: National Bureau of Economic Research, 1976.

Viscusi, W. Kip, "Union, Labor Market Structure, and the Welfare Implications of the Quality of Work," *Journal of Labor Research*, Spring 1980, *1*, 175–92.

The Lulling Effect: The Impact of Child-Resistant Packaging on Aspirin and Analgesic Ingestions

By W. KIP VISCUSI*

In 1972 the Food and Drug Administration imposed a protective bottlecap requirement on aspirin and other selected drugs. This regulation epitomizes the technological approach to social regulation. The strategy for reducing children's poisoning risks was to design caps that would make opening containers of hazardous substances more difficult. This engineering approach will be effective provided that children's exposure to hazardous products does not increase. If, however, parents leave protective caps off bottles because they are difficult to open, or increase children's access to these bottles because they are supposedly "child proof," the regulation may not have a beneficial effect.

Indeed, in this case there was no significant impact of the regulation on aspirin poisoning rates, but there has been an alarming, upward shift in the trend of analgesic ingestion rates since 1972. The source of this pattern appears to be attributable to a general reduction in parental caution with respect to such medicines, which has had an adverse spillover effect on unregulated products. The economic mechanisms involved can be best understood by considering the nature of individuals' response to regulatory protection.

I. The Lulling Effect: A Conceptual Analysis

One can distinguish three different mechanisms by which protective packaging requirements can lead to actions on the part of parents and their children that are at least potentially counterproductive. First, regulations will lead to a reduction in safety-related efforts for the affected product. Second, the regulation may produce misperceptions that lead consumers to reduce their safety precautions because they overestimate the product's safety. Finally, if there are indivisibilities in one's actions (for example, choosing whether to keep medicines in a bathroom cabinet or in the kitchen), regulating one product may affect the safety of other products. These effects are quite general and are not restricted to the case of protective bottlecaps.

The existing theoretical literature on individual responses to regulatory protection began with the analysis by Sam Peltzman (1975), who showed that seatbelts would lead to increased driving intensity (for example, less caution or higher speeds). The economic mechanism generating this effect is similar to that which produces adverse incentives or moral hazard problems in the insurance context. As one reduces either the probability of a loss or the size of the loss, individual incentives to take precautionary actions will be reduced. Regulations function much like insurance in this regard, with the only difference being that one need not pay an insurance premium. (There may, however, be an effect of the regulation on the product price.)

In my 1979 article, I derived a similar result for the case of worker safety for quite general classes of risk-averse preferences, where the safety measures also affected the wage rate. Except in the case of very stringently enforced government regulations, firms would not make technological changes in the workplace that were counterproductive. Compliance with policies such as seatbelt and bottlecap requirements is less discretionary, however, so one cannot rule out counter-

*Professor and Director, Center for Study of Business Regulation, Fuqua School of Business, Duke University, Durham, N.C. 27706. This research was supported by a grant from the National Science Foundation to the Center for Study of Business Regulation, Fuqua School of Business, Duke University. I am indebted to Robert Viscusi for insightful comments and to William Evans for able research assistance.

productive regulatory effects in these instances.

To investigate these effects more formally, consider a simple model that captures the essential features of these analyses. Let s be the stringency of the government policy and e be the precautionary effort, where each of these reduces the probability $p(e,s)$ of an accident at a diminishing rate. Alternatively, one can make the mechanism of influence the size of the accident loss L, as in Peltzman, but for purposes of this model I will make L a constant. The individual's effort e generates a disutility $V(e)$, where $V', V'' > 0$. Finally, let the person have an income level I.

The payoff in the case of an accident is $I - V(e) - L$, and the payoff if there is not an accident is $I - V(e)$. The individual's expected utility (assuming risk neutrality) is $I - V(e) - p(e,s)L$. In setting the optimal level of e, one equates the marginal reduction in the loss p_eL to the marginal value of the effort $-V_e$, leading to the optimal point A in Figure 1.

The effect of the regulation on safety effort will be negative, or

$$\frac{de}{ds} = \frac{-p_{es}}{p_{ee}L + V_{ee}} < 0,$$

provided the $p_{es} > 0$ (or $L_{es} > 0$ in Peltzman's loss model). For safety efforts to decline, the safety regulation must reduce the marginal safety benefits from precautionary efforts, that is, the reduction in the expected loss from higher levels of effort is less negative than before. One will then choose a point to the left of point B on the EL_1 curve in Figure 1. This effect should not be particularly controversial. Few would question the opposite relationship where individuals increase their precautions when moving from EL_1 to EL_0. For example, one will drive more carefully on icy streets and reduce cigarette smoking if exposed to synergistic asbestos risks.

What is more problematic is whether the reduction in precautions will be so great that there will be a reduction in safety to a point to the left of point C on EL_1. The conditions for this to occur have never been investigated and are quite stringent. It is not sufficient that the marginal expected loss reduction at point C be no greater than at A. The requirement is stronger since the marginal disutility of effort V_e will be lower at lower effort levels. To equate the marginal expected loss reduction p_eL to the marginal effort cost $-V_e$, the loss curve EL_1 must be flatter at point C than EL_0 was at point A. Since there is the additional restriction that EL_1 lie below EL_0, it will be difficult to meet these requirements.

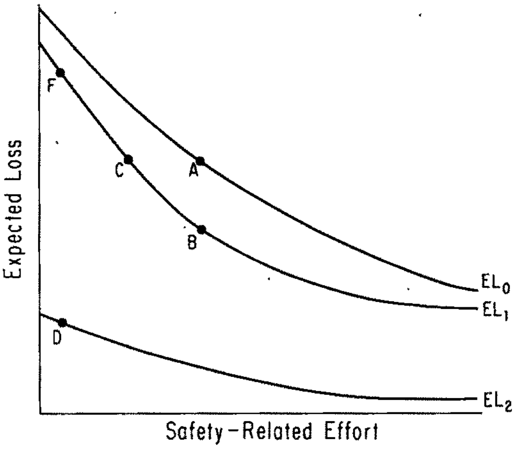


FIGURE 1. PRECAUTIONARY BEHAVIOR AND EXPECTED LOSSES

The chance that the impact of protective regulations may be counterproductive may be enhanced if individuals either do not perceive accurately the accident probabilities, or do not fully bear the accident costs. If parents assume "child-resistant" caps are child proof, they may overestimate the safety associated with these products. Similarly, since there is some evidence that individuals tend to set very small probabilities equal to zero, the safety-enhancing properties of caps may reduce the risk so much that parents ignore the poisoning risk. In each case, safety precautions will decline, perhaps to so great an extent that overall safety is reduced. For example, parents may select point D off of their perceived loss curve EL_2 in a situation where the true loss curve is EL_1 and the actual outcome is point F.

A similar effect could occur if parents do not fully value the welfare of their children, or if drivers do not fully internalize the accident costs to pedestrians and other parties. Unlike the case of biased probabilistic beliefs, this modification in the problem need not entail a shift in the relative values of the accident loss in the regulated and unregulated situations. Thus, the EL_0 and EL_1 curves may both simply shift proportionally. In contrast, misperceptions such as those discussed above necessarily lead to a comparatively greater downward shift in the perceived expected loss, increasing the chance of a counterproductive effect.

One's precautionary actions may affect the safety of unregulated products as well as those that are regulated. In the case of child-resistant bottlecaps, parents may make overall decisions regarding the storage of medicines. Should they keep all of the medicines in the bathroom cabinet, on a kitchen shelf, or in a safety-latched drawer? More generally, should they worry about access to medicines or undertake only a mild level of precautions, since the most hazardous products are presumably protected by child-resistant containers?

The analytics of this effort decision parallel that given above, where the only difference is that the EL_0 and EL_1 curves are a weighted average of the component risks, where some products are protected and others are not. A joint risk curve EL_1 will tend to shift downward less in response to a regulation than a comparable curve for a particular regulated product, since the presence of the unregulated product will dampen the response.

There is a clearcut empirical test of whether indivisible actions such as this play a role. If there are such spillover effects, the reduction in safety-enhancing efforts induced by the regulation should increase the risk posed by the unregulated product. In addition, the safety improvement of the regulated product will be reduced at least in part by the reduction in individual precautions. The net effect on safety could be adverse for the regulated product or for both products combined, but one must satisfy fairly stringent conditions for the net effect to be adverse unless misperceptions of the risk play a major role.

II. The Effect of Child-Resistant Bottlecaps

A widely touted product safety regulation success story is child-resistant bottlecaps. The first caps required under this regulation were for aspirin and selected drugs in 1972. Before the advent of protective packaging, manufacturers concentrated their efforts on measures such as decreasing the number of tablets per bottle, warning labels, and educational campaigns. Here I will summarize some of the results from my forthcoming study regarding the effectiveness of safety caps, which provides very strong evidence regarding the role of individual actions that differ in character from the seatbelt case (see Glenn Blomquist, forthcoming, for a review).

In 1971 aspirin was responsible for a fatal poisoning rate of 2.6 per million children under age 5, and by 1980 this rate had dropped to 0.6. The overall aspirin poisoning rate exhibited a similar drop, from 5.0 to 1.7 per 1,000. While these declines were dramatic, after taking into account the trend in aspirin poisonings and the decline in aspirin sales in the 1970's, there is no statistically significant impact of the regulation. This result was obtained using both a regulation dummy variable, which assumed a value of 1 in the 1972–80 period, and a variable that reflected the fraction of aspirin sold with safety cap bottles.

This fraction of capped bottles remained at just over half of all aspirin sold since firms were permitted to market one size of aspirin container without a child-resistant cap. Typically, firms chose the best selling size (the 100-tablet bottle).

Despite the constant sales share of safety capped aspirin, there has been a sharp increase in the proportion of aspirin-related poisonings associated with protective packaging. Whereas 40 percent of all aspirin poisonings in 1972 were from safety cap bottles, this figure rose to 73 percent by 1978. This pattern is noteworthy for two reasons. First, safety cap bottles are by no means risk free, as they account for a majority of the poisonings and a disproportionate amount compared to their sales. Almost half of all aspirin poisonings are from bottles that had been left open. Second, there appears to be an alarming increase in the rate of safety cap poisonings. While each of these effects may be attributable in part to consumers matching their aspirin bottle-type choice (with or without a cap) to whether or not they have young children, another factor may be that there is an increased degree of irresponsibility regarding medicines.

Such irresponsibility is consistent with evidence of an apparent spillover effect on previously unregulated analgesics, which include acetaminophen preparations such as Tylenol. Analgesic poisoning rates for children under age 5 escalated from 1.1 per 1,000 in 1971 to 1.5 per 1,000 in 1980. Even after taking into account increases in analgesic sales, 47 percent of this increase is attributable to an unexplained upward shift in the analgesic poisoning rate beginning in 1972. The coupling of the absence of any shift in the trend of aspirin poisoning rates with an upsurge in analgesic poisoning rates is consistent with the hypothesis that there is a significant indivisibility in safety precautions. Moreover, absence of a significant effect of safety caps on aspirin poisonings and the 47 percent unexplained shift in analgesic poisonings suggests that the impact of the regulation on balance was counterproductive, leading to 3,500 additional poisonings of children under age 5 annually from analgesics.

It is possible but unlikely that such a strong impact could emerge from fully rational consumer decisions. Moreover, this effect is not only large but reasonably widespread, as I have identified a similar pattern for prescription drugs, and for cleaning and polishing agents. A more likely explanation for these dramatic effects is that consumers have been lulled into a less-safety-conscious mode of behavior by the existence of safety caps. The presumed effectiveness of the technological solution may have induced increased parental irresponsibility.

A variety of regulatory efforts have sought to reduce individual risks through mandated technological changes. These measures will be effective if individual actions remain unchanged. In practice, these regulations will produce a lulling effect on consumer behavior because the perceived need for precautions will decline, potentially producing adverse spillover effects on the safety of other products. The strength of these impacts should highlight the importance of taking individual behavior into account when designing regulations intended to promote safety.

REFERENCES

Blomquist, Glenn, *Traffic Safety Regulation by NHTSA*, Washington: American Enterprise Institute, forthcoming.

Peltzman, Sam, "The Effects of Automobile Safety Regulation," *Journal of Political Economy*, August 1975, *83*, 677–725.

Viscusi, W. Kip, "The Impact of Occupational Safety and Health Regulation," *Bell Journal of Economics*, Spring 1979, *10*, 117–40.

______, "An Assessment of the Safety Impacts of Consumer Product Safety Regulations," Center for Study of Business Regulation Working Paper 83-10, Fuqua School of Business, Duke University, 1983.

______, *Regulating Product Safety*, Washington: American Enterprise Institute, forthcoming.

Automobile Safety Regulation and Offsetting Behavior: Some New Empirical Estimates

By ROBERT W. CRANDALL AND JOHN D. GRAHAM*

Suppose that engineers can demonstrate that air bags will reduce the risk of death in an automobile by 25 percent for any given frequency and severity of accidents. Would the installation of these devices necessarily reduce the fatality rate by 25 percent? The answer depends upon the response of drivers to the increased protection from dangerous accidents. If they increase their "driving intensity" (speed, recklessness, driving while intoxicated, driving in unsafe conditions, etc.), they may realize substantially less than a 25 percent reduction in expected fatalities. Such offsetting behavior is not irrational: it merely represents a substitution of the marginal benefits of driving intensity for the reduced marginal cost of risk.

If offsetting behavior actually occurs, it may be realized in increased risks for bicyclists, motorcyclists, and pedestrians. These externalities could be substantial unless there is a reduction in risk taking among these groups. As a result, the net effect of mandating air bags or any other safety device is far from obvious. There may be no net reduction in fatalities or serious injuries.

These theoretical considerations are at the core of Sam Peltzman's classic study (1975) of automobile safety regulation. For policymakers, however, the key question is *how much* offsetting behavior actually occurs. As Peltzman (1977) acknowledges, offsetting behavior could be trivial or substantial. In this paper, we explore this issue, providing new empirical estimates of the effects of crashworthiness standards established for automobiles over the past fifteen years. These standards have required the installation of lap-shoulder belts, energy-absorbing steering columns, head restraints, padded dashboards, crush-resistant passenger compartments, safer windshield mounting, more secure locks, and a variety of other features.

I. Previous Research

Peltzman (1975) used data from 1947 to 1965 to estimate the determinants of highway fatality rates. These results were then used to project fatality rates for the first seven years of federal safety regulation. These 1966–72 projections resulted in an overestimate of occupant deaths that was exactly matched by the underestimate in nonoccupant deaths, leading him to conclude that offsetting behavior had negated the potentially beneficial effects of the new safety standards.

A lively and emotional controversy followed. H. C. Joksch (1976), Paul MacAvoy (1976), and Richard Nelson (1976) criticized Peltzman's methodology but offered no new estimates. Peltzman (1976a) responded with incisive rebuttals. Leon Robertson (1977) produced different results with a revised model, but Peltzman (1977) and Glenn Blomquist (1981) argue that the revisions are *ad hoc* and inconsistent with a model of rational driver choice. Oscar Cantu (1980) has replicated Peltzman's findings with an independently constructed data base, but Graham and Steven Garber (1984) have shown that Peltzman's results are sensitive to plausible specification changes.

We cannot evaluate here the empirical literature which has accumulated on the offsetting driver-behavior issue, instead we simply report the results of two independent attempts to estimate the net lifesaving effects of federal automobile safety regulation. These results are new in their approach to simul-

*Senior Fellow, The Brookings Institution, Washington, D.C. 20036, and Visiting Lecturer and Research Fellow, School of Public Health, Harvard University, Boston, MA 02115.

taneity, the explicit measurement of automobile safety, and the comprehensive use of all data available through 1981.

II. New Empirical Estimates

Driving intensity is a complex behavioral concept that cannot be measured easily, even though there is an enormous amount of data on automobile usage. Like Peltzman, we assume that average vehicle speed is at least a proxy for such risk taking. We depart from Peltzman's approach, however, by specifying a simultaneous equation model of automobile safety:

$$(1) \qquad DR = f(S, K, A, H, V, R),$$

$$(2) \qquad S = g(DR, K, A, Y, P),$$

where DR is the relevant highway death rate per vehicle mile; S is average vehicle speed; K is a measure of driving by "kids" (high-risk youth); A is per capita alcohol consumption, a proxy for alcohol-impaired driving; H is a vector of attributes describing highway design; V is an index of the average weight of the vehicle fleet; R is a proxy for the degree of crashworthiness required by federal regulation; Y is the value of a driver's time (his earned income); and P is an index of the cost of an accident.

The net effect of R on DR may be estimated consistently by applying ordinary least squares to the reduced-form, death-rate equation:

$$(3) \qquad DR = h(K, A, H, V, R, Y, P).$$

A more elaborate simultaneous model might treat K, A, and V as endogenous variables. In this paper, however, we only report estimates of equation (3). This is the first attempt to build a simultaneous model of highway safety, as was originally suggested by MacAvoy in his critique of Peltzman's model.

If driving intensity is increased by safety regulation, then R should reduce fatality rates for passenger car occupants by an amount less than that predicted by safety engineers (15 to 35 percent). And R will increase death rates for those not in passenger cars. But how can R be measured given the lack of sales data on cars with specific safety features? We provide estimates using two alternative measures—the proportion of miles driven by cars built since federal automobile safety regulation began in 1968 (R_1) and a weighted measure of such miles where the weights reflect GAO (1976) estimates of improved occupant protection built into successive post-1965 model cars (R_2). The first measure of vehicle safety, R_1, rises from 0 to 1.0 as pre-1968 cars are scrapped. The second measure, R_2, declines from 1.0 to 0.77 as all pre-1967 cars leave the vehicle stock. The latter measure is based on a GAO study of occupant survival rates in North Carolina crashes involving cars of model years 1966–74. The results of Graham's (1983) time-series analysis appear in Table 1, while the Crandall (1983) results are reported in Table 2.

The first two equations from Graham's work are linear reduced-form equations that include a dummy variable for imposition of the 55 mph speed limit (*MAX55*), a measure of the share of miles operated on rural roads (*RURAL*), the share of miles driven on limited access highways (*LACCESS*), and the average weight of cars on the road (V). In addition, a time trend is included. These equations are linear in all variables.

TABLE 1—GRAHAM'S TIME-SERIES ESTIMATES OF THE EFFECT OF SAFETY REGULATION ON THE OCCUPANT AND PEDESTRIAN DEATH RATE, 1947–81

Independent Variables	Passenger Car Occupant Death Rate	Pedestrian Death Rate
R_1 (Safety)	-1.479^a	0.259
K (Youth)	20.52	12.56^a
A (Alcohol)	2.472^a	0.754^a
RURAL	8.994^a	−2.911
LACCESS	−4.741	−0.935
MAX55	0.122	−0.005
Y (Income)	1.032^a	0.004
P (Cost)	−1.915	-1.587^a
TREND	-0.115^a	-0.083^a
V (Weight)	-0.339^a	−0.128
$\bar{R}^2$	0.981	0.984

[a]Statistically significant at the 5 percent confidence level.

TABLE 2—CRANDALL'S TIME-SERIES ESTIMATES OF THE EFFECT OF REGULATION ON OCCUPANT AND PEDESTRIAN DEATHS, 1947–81

Independent Variables	Passenger Car Occupants	Pedestrians and Bicyclists
R_2 (Safety)	2.331[a]	−0.8010
K (Youth)	0.7325	0.2632
A (Alcohol)	0.2941	0.4624
RURAL	0.7506[a]	−0.1407
LACCESS	−0.0653	0.0801[a]
TRUCK	0.3893[a]	–
Y (Income)	1.062[a]	0.9249[a]
TREND	−0.0259[a]	−0.0413[a]
V (Weight)	−2.860[a]	−4.483[a]
MILES	1.367[a]	−0.4781
$\bar{R}^2$	0.979	0.995

[a] See Table 1.

The Crandall results are log linear in form and the dependent variables are occupant and pedestrian deaths, not the death rates, since vehicle miles (*MILES*) is introduced as an independent variable. The occupant death equation contains an explicit variable for the share of truck miles relative to total annual vehicle miles (*TRUCK*).

In both sets of estimates, the regulation proxies perform spectacularly as regressors in the passenger car occupant death (rate) equations. Since R_1 rises with increased safety towards 1.0, its coefficient is expected to be negative in the occupant equation. Similarly, since R_2 declines towards 0.77 with increasing safety, its coefficients should be positive. In both occupant equations, the regulatory variables assume coefficients that are larger than might be expected by the GAO study. Cars appear to be getting even safer for occupants than the 23 percent estimate of improvement through 1971 would suggest. The estimate for R_1 is, however, within the range of improvement predicted by more optimistic engineering studies reviewed in Graham.

Offsetting behavior is not apparent in the Graham estimate of the pedestrian death rate, but it is present in the Crandall estimate at the 10 percent confidence level. However, even the latter estimate is much smaller than that required for an increase in pedestrian-cyclist deaths to fully offset the decline in occupant deaths.

III. Conclusions

The time-series evidence reported here reveals some offsetting behavior, but the intrinsic engineering effects of safety devices appear to swamp the behavioral responses. That result differs substantially from Peltzman's original empirical result. As a check, we are also producing estimates from cross-sectional analyses of state and car-vintage data. In addition, a variety of other investigators are studying directly the behavior of drivers before and after imposition of compulsory seat belt usage laws in Canada and England. Results from these studies should clarify further the empirical significance of the economist's behavioral response prediction.

REFERENCES

Blomquist, Glenn, "Traffic Safety Policy: An Economic Evaluation," Working Paper, College of Business and Economics, University of Kentucky, October 1981.

Cantu, Oscar R., "An Updated Regression Analysis on the Effects of the Regulation of Automobile Safety," Working Paper No. 15, School of Organization and Management, Yale University, 1980.

Crandall, Robert, "The Effects of Regulation on Automobile Safety," manuscript, The Brookings Institution, 1983.

Graham, John D., "Automobile Safety: An Investigation of Occupant Protection Policies," unpublished doctoral dissertation, Carnegie-Mellon University, 1983.

______ **and Garber, Steven,** "Evaluating the Effects of Automobile Safety Regulation," *Journal of Policy Analysis and Management*, forthcoming 1984.

Joksch, H. C., "Critique of Sam Peltzman's Study," *Accident Analysis and Prevention*, No. 2, 1976, *8*, 129–137.

MacAvoy, Paul, "The Regulation of Accidents," in H. G. Manne and R. L. Miller, eds., *Auto Safety Regulation: The Cure or the Problem*, Glen Ridge: Thomas Horton & Daughters, 1976, 83–88.

Nelson, Richard R., "Comments on Peltzman's Paper on Automobile Safety Regulation," in H. G. Manne and R. L. Miller, eds., *Auto Safety Regulation: The Cure or the Problem*, Glen Ridge: Thomas Horton & Daughters, 1976, 63–72.

Peltzman, Sam, "The Effects of Automobile Safety Regulation," *Journal of Political Economy*, August 1975, *83*, 677–725.

______, (1976a) "The Effects of Automobile Safety Regulation: A Reply," *Accident Analysis and Prevention*, No. 2, 1976, *8*, 139–142.

______, (1976b) "The Regulation of Automobile Safety," in H. G. Manne and R. L. Miller, eds., *Auto Safety Regulation: The Cure or the Problem?*, Glen Ridge: Thomas Horton & Daughters, 1976, 1–51.

______, "A Reply to Robertson," *Journal of Economic Issues*, September 1977, *11*, 672–678.

Robertson, Leon S., "A Critical Analysis of Peltzman's 'The Effects of Automobile Safety Regulation'," *Journal of Economics Issues*, September 1977, *11*, 587–600.

U.S. Government Accounting Office, *Effectiveness, Benefits, and Costs of Federal Safety Standards for Passenger Car Occupants*, Report of the Comptroller General of the United States, Washington, July 7, 1976.

THE THEORY OF CONTRACTS AND AGENCY: IMPLICATIONS FOR THE BEHAVIOR OF WAGES

Contract Costs and Administered Prices: An Economic Theory of Rigid Wages

By BENJAMIN KLEIN*

Macroeconomists have long puzzled over the fact that nominal wages are largely insensitive to aggregate economic activity. Until relatively recently this phenomenon and the supposedly resulting unemployment was "explained" by the Keynesian assumption of predetermined money wages. Although this view can still be found in most textbooks, a new theoretical view of the labor market has developed which attempts to explain this phenomenon by emphasizing the fact that most labor market relationships are de facto long term, and that workers are risk averse.[1] A labor contract is considered similar to a mortgage with the wage merely an installment payment on a long-term "implicit" commitment to transfer a certain amount of wealth in exchange for a certain amount of labor services (see Robert Hall, 1980). Under such circumstances we would expect the time path of wage payments to be determined solely by the convenience of the transacting parties. Since workers are risk averse, and are not likely to be able to borrow or lend as cheaply as firms, the firm pays a wage over time that smooths out worker income fluctuations.

While risk aversion may exist in the labor market, rigid wages are an unlikely substitute for worker savings. These contracts generally do not cover the poorest workers nor do they smooth real (as opposed to nominal) wages. Most importantly, rigid or administered prices appear to be present in many markets where large corporations are on both sides of the transaction, and hence where risk aversion is unlikely to be the prime concern. This paper attempts to apply the theoretical insights that can be obtained from these corporate contractual arrangements to the labor market and thereby begin to develop a microeconomic foundation for macroeconomic analysis based on the assumption of risk neutrality.

I. The "Hold Up" Problem

An economist who asks why wages are sticky is essentially asking why labor is not sold in a spot auction market. Restated in this way, the answer to the question is fairly obvious. Labor (and most other inputs) are purchased by explicit and implicit long-term contracts rather than in spot markets because of the presence of firm-specific investments. My earlier work with Robert Crawford and Armen Alchian (1978) analyzed the potential "hold up" problem involved when such specific investments are made by one of the parties to a transaction. After a firm invests in an asset with a low salvage value and a quasi-rent stream highly dependent upon some other asset, the owner of the other asset has the potential to hold up by appropriating the quasi-rent stream. For example, one would not build a house on land rented for a short term. After the rental agreement expires, the landowner could raise the rental price to reflect the costs of moving the house to another lot.

*University of California, Los Angeles, CA 90024. I am grateful to Roy W. Kenney and Kevin M. Murphy for valuable comments and to The Sloan Foundation grant to UCLA for the study of contractual relationships for research support.

[1] Important theoretical contributions have been made by Martin Baily (1974), Costas Azariadis (1975), and Donald Gordon (1974). For contrary views in the spirit of my analysis, see Michael Wachter and Oliver Williamson (1978) and David Mayers and Richard Thaler (1979).

In the labor market, these considerations are paramount. Many jobs require significant firm-specific investments, including the investment of time on the part of the worker in learning to work with the specific team of workers within the firm's specific organizational framework.

Because the firm's brand name in most cases is likely to be relatively larger than the worker's, the worker can be expected to make much of the specific investment and the firm guarantee that it will not hold up the worker by reducing his wage below the value of his marginal product. A firm generally has lower costs of creating brand name capital and hence contract fulfillment credibility because of its increased repeat-purchase frequency. While a firm is always hiring additional workers and must bear the future cost from cheating now, workers have limited lifetimes and working opportunities. In addition, because of the larger size of firms compared to individual workers, cheating firms are likely to become known more quickly than cheating workers, reducing the short-run cheating potential for firms relative to workers.

However, it is unlikely that the equilibrium will have the worker making the entire specific investment. Letting the firm finance some of the specific investment reduces the required firm brand name capital while not significantly increasing the hold up potential on the part of the worker. As long as the worker continues to make a significant investment, his threat to leave unless the wage is adjusted upward is not credible. Hence financing of the firm-specific human capital investment is likely to be shared to some extent by the worker and the firm.

II. Explicit and Implicit Contractual Solutions

Although I have emphasized vertical integration as a mechanism to solve the hold up problem, it is not possible in the labor market. However, vertical integration need not be relied on if, prior to investment, one can write a complete enforceable long-term contract. It may not be necessary, for example, to own the land upon which one intends to build a house if an enforceable long-term lease is obtained before the house is built. The interesting economic question relates to the type of long-term contract that is likely to be most efficient. In particular, while the presence of specific human capital implies the necessity for a long-term contractual relationship, why are wages often set by a long-term implicit contract rather than by a long-term explicit contract?

The existing literature makes no economic distinction between explicit and implicit contracts. Contracts are referred to as "implicit" solely in the sense of an unwritten understanding. One does not observe an explicit, written contract, yet the transacting parties are assumed to behave as if it existed. But this is a distinction without a difference. All contracts, whether written or not, are assumed to be costlessly enforceable.

It is useful to classify contracts or elements of contractual relationships by the enforcement mechanism adopted by the contracting parties. Contractual performance can be assured either by explicit sanctions which are imposed by a third party (say, a court), or by implicit two-party sanctions, namely termination of the contractual relationship. If the individual facing termination expects to be earning a quasi-rent stream in the future, the present discounted value of which is greater than the immediate short-run gain from breach divided by the probability of detection, the threat of termination will be sufficient to assure performance.[2]

Court-enforced sanctions have the advantage over two-party sanctions in that money can be awarded to one or the other party *ex post*, and hence the timing of performance by the transacting parties is irrelevant. Timing is, on the other hand, crucial in the implicit contract case. Since the only sanction is the termination of the agreement and everyone keeps what they have at the point of termination, it is crucial that the future expected premium stream be greater than the hold up potential at every point in time. With a court-imposed sanction, on the other hand, the transacting parties can agree to do things in the future that will not *ex post* be incentive compatible. The court can,

[2] See my 1980 paper for a discussion of this mechanism in the franchising context.

in a sense, put things back together again. Hence exchanges can be structured so that performance by the parties need not be simultaneous. Performance can be sequential without an expected future premium stream being present in the correct magnitude at every point in time.

In choosing a contractual arrangement, transactors will trade off the costs of enforcing performance via these alternative mechanisms. Explicit contracts entail the transaction costs of writing everything down. These costs refer not merely to the ink costs involved, but, in an uncertain world with a large number of possible contingencies, to the significant real resource costs of discovering all the possible things that can happen in the future, and figuring out the optimal response by the transacting parties for all these hypothetical, largely irrelevant, states. Individuals will also devote time and money in attempting to obtain an informational advantage over their transacting partners, and in bargaining over mutually acceptable contingent terms.

Explicit contracts are also costly to enforce because particular performance, such as the level of energy an employee is to devote to a complex task, may be prohibitively costly to measure and hence to specify contractually. Therefore contractual breach and the extent of damages will be difficult to prove to the satisfaction of a third-party enforcer such as a court. Transacting parties generally rely on some proxy measure of performance, but even these proxies are often extremely complex. An employer may observe and monitor many aspects of employee behavior before deciding on termination or promotion, and these signals may be extremely costly to communicate to the court.

It is therefore unlikely to find in the real world, as opposed to the standard economic model, complete, fully contingent, court-enforced contracts. Such contracts are not cheaply specifiable nor cheaply enforceable. All contracts are, by necessity, somewhat vague. However, it is also highly unlikely to find a real world corner solution in the other direction. While incomplete two-party arrangements economize on the transaction costs of writing and enforcing complete explicit contracts, they entail the costs of performance-assurance premiums and the possibility that inefficiently large "brand name" (firm specific, nonsalvageable) investments will have to be made.

Most actual contractual arrangements, including labor contracts, can be expected to consist of a combination of explicit and implicit enforcement mechanisms. Some elements of performance will be specified and enforced by third-party sanctions, while the residual elements of performance will be enforced by the implicit threat of termination of the transactional relationship. The future expected quasi-rent stream received by the worker on his specific investment will generally be more than sufficient to prevent shirking, and the firm's brand name capital will then prevent the firm from holding up the worker for the specific investment above this minimum amount.

III. Explicit Contract Rigidity

Another cost of explicit contracts compared to implicit contracts is the increased rigidity of such arrangements. This implies that an explicit contract term, such as price, is more likely to differ from the "perfectly competitive" level. This results in costly resource misallocations, which may be avoided by more flexible implicit contract terms.

Consider a particular real world example —the supply of automobile bodies by Fisher Body Corporation to General Motors.[3] In 1919, as the production process for automobiles was shifting from individually constructed open, largely wooden, bodies to metal closed-body construction, General Motors entered a contractual agreement with Fisher Body for the supply of closed-auto bodies. Since Fisher Body had to make a highly specific investment in stamping machines and dies, it is obvious that a short-term spot contract could not be used. Instead, a long-term (ten-year) fixed formula price con-

[3] The manufacturing agreement between GM and Fisher Body can be found in the minutes of the Board of Directors of Fisher Body Corporation for November 7, 1919.

tract was negotiated with the price set equal to cost plus 17.6 percent.

However, even if price is effectively fixed, a buyer may be able to hold up a seller who has made a buyer-specific investment by threatening, unless some price adjustment or side payment is made, to vary quantity demanded, including the threat of complete termination. To prevent this, the General Motors-Fisher Body contract included an exclusive dealing clause, whereby GM agreed to buy over the period of the contract all their closed bodies from Fisher. This arrangement significantly reduced the possibility of GM acting opportunistically after Fisher made the specific investment in production capacity.[4]

Labor contracts often include lay off terms which are analytically similar to this exclusive dealing arrangement adopted by GM and Fisher. If, for example, in the face of a claim of declining demand, a firm must keep wages fixed and lay off workers, and is prevented from hiring additional workers of the same type at a lower wage, a contractual arrangement exists which substantially reduces the incentive for the firm to claim opportunistically a false decrease in demand. This seniority-type rule implies that the firm must also hurt itself when it threatens to layoff workers.

Within this exclusive dealing-fixed wage context, the firm may still attempt to hold up workers by varying quantity demanded. Since the firm may not be hurt as much as the worker who made the specific investment, it may be able to credibly threaten layoffs to appropriate the worker's quasi rents. To prevent this, the contract may require payment whether or not workers are working.[5]

While these contractual arrangements may effectively prevent the hold up, they may produce severe misallocation problems because of their preset price terms. As noted above, because of information and measurement costs, it is extremely difficult, if not impossible, to specify all performance terms ex ante. In the GM-Fisher case, omissions in the contract were glaring and caused problems almost immediately. First, although the price was set on a cost-plus basis, cost was defined exclusive of interest on invested capital. Given the absence of a capital cost pass through, Fisher shifted towards a low-capital-intensity form of production with resulting higher prices to General Motors. In addition, because transportation costs were reimbursable as part of the price formula, Fisher refused to locate their body plants adjacent to GM's assembly plant, a move which GM claimed was necessary for production efficiency.[6]

These difficulties were not entirely unanticipated by General Motors and Fisher Body. In an attempt to prevent such problems, the contract included provisions that the price charged GM could not be greater than what Fisher charged other automobile manufacturers for similar bodies. However, this "most favored nation" clause proved to be ineffective, apparently because of the difficulty of defining what is "similar."[7]

It is common for transacting parties, including participants in the labor market, to use such "price protection" rules to prevent the hold up. In this way a price increase or decrease to any supplier is guaranteed to be given to all suppliers. Established workers that are "locked in" by a specific investment

[4] If it is efficient for the buyer to purchase from many sources, the contract may call for the buyer to purchase all that an individual seller can supply at a preset price. Natural gas supply contracts made with monopolistic pipeline companies are an obvious example. Analogously, if a buyer makes a seller-specific investment, an agreement to supply the buyers "requirements" will effectively prevent the seller hold up.

[5] See Martin Feldstein (1976). Fixed-price take-or-pay contract terms made by natural gas pipeline companies are an obvious example of this in a nonlabor market. The recent problems experienced with these contracts as market prices have declined drastically are illustrative of our concerns.

[6] See deposition and direct testimony of Alfred P. Sloan, Jr. in *United States v. DuPont & Co.*, 366 U.S. 316 (1961), 186–90 (April 28, 1952), and 2908–14 (March 17, 1953).

[7] The original contract also stated that the price could not be greater than the average market price of similar bodies produced by companies other than Fisher. In addition, it included provisions for compulsory arbitration in the event of any dispute regarding price. These provisions also proved ineffective.

are protected by the necessity of the firm to hire new workers. While such clauses may appear to be collusive and to produce rigidity, they efficiently raise the cost to the firm of cheating and thereby lower the firm's required brand name capital.

These difficulties experienced by GM are inherent to any long-term explicit fixed-price contract. They are produced by rigid contract terms, such as price determined by a preset formula, and not necessarily by fixed prices. Prices are "sticky" in the sense that they do not track market conditions perfectly, that is, in the sense that formulas are imperfect. A firm may attempt to index worker compensation, for example, to market conditions, but this will necessarily be imperfect when attempting to track the return on a firm-specific human capital investment. Changes in economywide indices such as "the" price level will move in the incorrect direction in response to relative (firm or industry) shocks. They therefore will be used only when the economywide variance is large.

The benefit from such imperfect explicit contractual arrangements is obvious. They may prevent in an inexpensive manner a hold up by the firm. But, as in the GM-Fisher case, the cost is also obvious. Sellers or buyers may take advantage of inappropriate prices and resources will be misallocated in the process. The question is whether these inefficiencies are small enough to more than compensate for the hold up prevention benefit of fixing terms. Often, especially when demand and/or supply changes are significantly greater than anticipated, the inefficient under- or overutilization of fixed-price inputs becomes intolerable. In addition to the distribution effects of an incorrect price and the possible bankruptcy of one of the parties, real resource misallocation costs are created by the supplier (demander) attempting to take advantage of the high (low) price. Further, the transactor placed at a disadvantage may attempt to renege on the contract, creating unnecessary disruptions and legal expenses.

What do these inefficiencies consist of in the labor market? Given the presence of firm-specific capital, an incorrect wage is unlikely to lead to worker termination of the firm. In addition, if the possibility of firm termination of the worker exists, worker shirking is not likely to occur. The wage would have to fall to the point where the discounted value of the premium above the worker's opportunity wage is less than the short-run gain from shirking. The inefficiencies relate to the fact that the worker can stop making continuing firm-specific investments. This is analytically equivalent to partial worker termination of the firm and will be costly to the firm given its complementary investments.

IV. Implicit Contract Flexibility

The GM-Fisher Body contractual difficulties led in 1926 to their merger, but vertical integration is not the common solution adopted by transacting parties in such circumstances. Transactors rely, instead, on each other's brand names, namely the present discounted value of quasi rents connected with the transaction, to provide adjustments in the unspecified terms of a contract. In the GM-Fisher Body case, the current contractual period (ten-year) demand drew unanticipatedly rapidly relative to the future demand so that the loss of future rents to Fisher from the failure of GM to renew the contract became insufficient to assure implicitly understood performance.[8] The "short-run" Fisher cheating potential became greater than anticipated at the time the contract was made and the arrangement broke down.

More generally, contractual adjustments and renegotiation to recognized changing conditions will occur to prevent such a collapse. This is the advantage of implicitly setting some contractual terms rather than attempting to explicitly fix all contractual terms *ex ante*. While the brand name costs (firm-specific premium rents) associated with implicit contractual enforcement can be

[8] From 1918, when closed bodies were essentially a novelty, demand grew by 1924 to account for more than 65 percent of GM automobile production. See *Sixteenth Annual Report*, General Motors Corporation, year-ended December 31, 1924.

saved if contract terms are explicitly set, implicit contracts can be freely terminated and therefore the transactors "have an out" if market conditions change unexpectedly. Flexibility in price is enforced by the threatened loss of future rents from termination. As long as sufficient brand name capital exists, contract terms will adjust to all market changes that both parties are informed about.

If, however, only one party to the transaction (say, the buyer) is aware of the changing conditions that necessitates a price change, he may decide to keep price unchanged so as to minimize seller-monitoring expenditures and inefficient adjustments. The asymmetric information may be, for example, that the firm knows that his demand and the worker's value of marginal product has declined without the worker knowing that this has occurred. Even though the firm is operating under an implicit contract which permits a change in the wage, in such a situation he may decide not to make the adjustment because such a change may tend to create suspicion on the part of the worker regarding the purpose of the contract alteration. The worker may believe that this is merely an attempt by the firm to seize some of his firm-specific rents. This in turn will lead the worker to increase his costly monitoring activities and, if he believes the change is unjustified, reduce his continuing investment in firm-specific human capital. Therefore the firm may optimally keep the wage unchanged, foregoing the benefit of a correct marginal price, but preventing the cost of inappropriate worker-investment decisions.

Firms with extremely large brand names, and hence much to lose from cheating workers, are less likely to take the chance of cheating. Because worker estimates of the probability of a firm cheating in such a case are so low, workers are unlikely to respond to wage decreases by inefficiently reducing firm-specific investments. Therefore it will be optimal for such firms to freely adjust wages. This may explain why Japanese firms that possess such large brand name capital (because of their very high anticipated growth rates and hence the high associated future costs of currently being detected cheating) have such flexible wages.

V. Conclusion

If we are to explain satisfactorily the form of particular complex contracts adopted in the marketplace, we must consider the cost of enforcing performance in the particular transaction under investigation. As a useful starting point of analysis, I have outlined a general theoretical framework of contract enforcement. The important economic questions examined within this framework relate to (a) how incomplete the contract is likely to be, that is, how much reliance will be placed on implicit rather than explicit enforcement mechanisms, (b) what explicit terms are likely to be used in the contract, and (c) what responses to unexpected changes are likely to be made by the transacting parties.

As I have shown, the extent of reliance on implicit enforcement mechanisms is dependent in part upon the costs of creating brand name capital compared to the misallocation costs of incorrect explicitly fixed-contract terms. Particular explicit contractual terms, such as exclusive dealing and price protection provisions, may appear noncompetitive, but they efficiently economize on the required brand name capital without imposing too high a misallocation cost. A major benefit of an implicit contract is that the transacting parties can adjust to symmetric information that is not written, *ex ante*. The adjustment response will depend upon the magnitude of the distortion present and the quantity of brand name capital that exists.

A labor contract, although it is a long-term commitment, is not like a mortgage contract. While the amount of wealth to be transferred to the worker over the life of the contract would be reasonably specified in an enforceable way, the supply of labor services to be transferred to the firm over the contract life cannot be so specified. Risk-neutral firms will not commit themselves to an explicitly long-term, fixed-price relationship not because of the potential distribution effects caused by unanticipated changes in the market wage, but because they want to have "an out" so that contract terms can be adjusted and workers can be terminated.

It is important to recognize that the contractual terms and responses I have examined imply price rigidity only when compared to the unrealistic spot market alternative of the standard economic paradigm. Given the presence of firm-specific capital, and hence a potential hold up, the long-term contracts I have been investigating are *more* flexible than the relevant alternative benchmark of a long-term fixed-price contract.

Finally, I should note that all of this has not brought us far in our attempt to understand macroeconomic fluctuations. While sticky wages would produce unemployment within the context of a spot auction market, there is no apparent reason for such a response within our framework. Firms can be expected to possess sufficient brand name capital to hire the correct amount of labor independent of the short-run behavior of wages. Further theoretical analysis and much needed empirical work is required before we can solve this puzzle.

REFERENCES

Azariadis, Costas, "Implicit Contracts and Underemployment Equilibria," *Journal of Political Economy*, December 1975, *83*, 1183–1202.

Baily, Martin Neal, "Wages and Employment Under Uncertain Demand," *Review of Economic Studies*, January 1974, *41*, 37–50.

Feldstein, Martin, "Temporary Layoffs in the Theory of Unemployment," *Journal of Political Economy*, October 1976, *84*, 937–57.

Gordon, Donald F., "A Neo-Classical Theory of Keynesian Unemployment," *Economic Inquiry*, December 1974, *12*, 431–59.

Hall, Robert E., "Employment Fluctuations and Wage Rigidity," *Brookings Papers on Economic Activity*, 1:1980, 91–123.

Klein, Benjamin, "Transaction Cost Determinants of 'Unfair' Contractual Arrangements," *American Economic Review Proceedings*, May 1980, *70*, 356–62.

______, **Crawford, Robert G. and Alchian, Armen A.,** "Vertical Integration, Appropriable Rents and Competitive Contracting Process," *Journal of Law and Economics*, October 1978, *21*, 297–326.

Mayers, David and Thaler, Richard, "Sticky Wages and Implicit Contracts: A Transactional Approach," *Economic Inquiry*, October 1979, *17*, 55–74.

Wachter, Michael L. and Williamson, Oliver E., "Obligational Markets and the Mechanics of Inflation," *Bell Journal of Economics*, Autumn 1978, *9*, 549–71.

Incentives and Wage Rigidity

By EDWARD P. LAZEAR*

The notion that compensation may be structured to affect worker productivity is not new. Traditionally, piece rates have been the most common form of incentive pay. Recently, more elaborate bonus systems have begun to creep into the American labor scene. Concurrent with the growth of creative compensation practices has been the development of a literature that describes the incentive effects which are associated with existing or hypothetical payment schemes. (See, for example, Bengt Holmstrom, 1982, and the references cited therein.)

A similar but distinct literature has considered wage rigidity that results when labor contracts are used as an alternative to a spot market. Most of the work centers around the idea that workers want to insure themselves against a variable wage stream. Others concentrate on the nature of contracts when information is asymmetric—either the worker or firm (or both) has information to which the other party is not privy. (See, for example, Robert Hall's and my 1984 article and the citations therein.)

This essay examines incentive arrangements to determine whether they contribute wage rigidity to an economy. Specifically, the attempt by employers to induce workers to produce efficiently may change the variability of wages over the business cycle, life cycle, and across individuals. International comparisons have revealed differences in wage flexibility across countries.[1] Can these differences be explained by the extent to which the countries use incentive compensation? Or, turning it around, are measured "business cycle" variations in wages mere reflections of worker incentive schemes?

The conclusion is that there is no simple relation of incentives to wage flexibility. Some incentive contracts add wage variance and reduce inflexibility. Others have the opposite effect. After rigidity is defined, the approach is to "prove by counterexample," discussing a number of incentive schemes. Although the list is by no means exhaustive, it covers most of the important ones and provides enough variety to demonstrate that there exists no clear link between incentive provision and wage rigidity.

I. Definition of Rigid Wages

In order to obtain a concise definition of rigid wages, it is useful to start with a simple model. Consider the simplest technology where output, q, is the sum of effort, u, and luck, e:

$$(1) \qquad q = u + e.$$

Assume initially that the variance of e is zero so that luck is not a factor. This is relaxed below.

Output sells at price V so that the firm's profit function is Profit $= Vq - Y$, where Y is the compensation of labor, the only factor of production. In a world of competitive factor and product markets, the zero profit constraint must hold so that

$$(2) \qquad Y = Vq.$$

Wage rigidity is defined relative to what would occur in a spot market. There are two variables that are of interest. Changes in V may occur over business cycles, or may reflect secular effects on the value of output. Changes in u can be thought of as differences across individuals with respect to work efficiency or distaste for effort exertion.

*University of Chicago, 1101 East 58th Street, Chicago, IL 60637, and National Bureau of Economic Research. Financial support was provided by the National Science Foundation.

[1] Robert Gordon (1982) found that the variability of wages in Japan exceeds that in the United States. Although little evidence exists, much has been made of the widespread nature of the Japanese bonus system.

In a spot market with full information, equations (1) and (2) imply that

$$\text{(3a)} \qquad \partial Y/\partial V = u + V(\partial u/\partial V),$$

$$\text{(3b)} \qquad \partial Y/\partial u = V.$$

Equations (3a) and (3b) serve as the criterion against which results are compared to determine wage rigidity.

II. Fixed Effort and Variable Effort

Before considering any specific incentive scheme, it is useful to point out that the ability to vary effort contributes income variation to an economy, even if workers are homogeneous. To see this, examine (3a). There are two terms on the right-hand side. When the price of output changes, income changes because each unit of effort is now more valuable (captured by the u term), but also because the optimal level of effort changes (reflected in the $V\partial u/\partial V$ term). Appropriately, wages are more variable in a world where effort is not supplied perfectly elastically.

Let us not take as given that workers adjust effort appropriately. After all, if no restrictions were placed on workers, and if income Y were totally independent of the level of effort, workers would choose to perform at the lowest possible level. The worker wants to choose u so as to maximize utility, assumed to be given as

$$\text{(4a)} \qquad \underset{u}{\text{Max}}\, Y(u) - C(u),$$

where $Y(u)$ is the income function that he faces (it may depend directly on q and only indirectly on u) and where $C(u)$ is the cost of effort function. The first-order condition for an optimum is the standard

$$\text{(4b)} \qquad Y'(u) = C'(u).$$

If q can be observed perfectly, then the first best solution can be achieved by paying a "piece rate," that is, letting $Y = Vq = Vu$. The pure piece rate is virtually synonymous with a spot market in this context, so it is not surprising that differentiation of the piece rate income function duplicates equations (3a) and (3b).

III. Imperfect Observability of Output

Few production environments lend themselves to costless and perfect measurement. Although the output of a salesman is measured easily, that of a vice president of finance is not. Firms often adapt to these difficulties by using some kind of sampling mechanism that requires only a periodic check of part of the worker's output. Without attempting to write down the optimal sampling rule, let us merely state that one possible incentive device is to sample the worker's output with some (optimally chosen) probability p. If the worker's output is not audited, then he receives one wage, $W(V)$, specified in advance and potentially a function of V. If the worker is audited, then he is paid $S(q, V)$.

It is trivial to show that a first best wage scheme is to have $W(V) = 0$ and $S(V, q) = Vq/p$ so that $Y(u) = p\ Vq/p = Vq = Vu$. An incentive scheme of this sort introduces cross-sectional wage variability where none would exist were output observed costlessly, or were effort levels given exogenously. Suppose that all individuals have the same $C(u)$ function. Equation (4b) ensures that they all select the same level of u. But incomes will vary: $1-p$ of the individuals receive $W(V) = 0$ in income and p of the individuals receive Vu/p. If a piece rate with 100 percent sampling were employed, then all workers would receive Vu. So additional wage variation is a result. ($W(V) = 0$ should be interpreted as the amount produced at the minimum observable effort level.)

Over the business cycle, changes in the price of the product are reflected perfectly in the average wage across workers. But $(1-p)$ of the workers who receive $W(V) = 0$ find that their wages are independent of the changes in product price, whereas p of the workers find that their wage is especially sensitive to changes in product price. Compare $\partial Y/\partial V = (1/p)(u + V\,\partial u/\partial v)$ with equation (3a). If p is small, most workers find that their wages are much more rigid than they would be in a world of exogenous effort with perfect information.

The additional variance that is associated with this particular incentive scheme is not a necessary consequence of either imperfect observability or of the desire to use an incentive compensation structure. For example, in this simple case, there is another first best compensation scheme that mimics a perfect information spot market exactly. Suppose that the worker is told that he will receive Vu if $u \geq u^*$, but will be penalized some amount x if it is detected that u has fallen below u^*. For any given p, there is a sufficiently large x such that workers always choose to produce at $u = u^*$. Under these circumstances, each worker's wage is always Vu, no cross-sectional variation in wages is introduced by this incentive plan, and (3a) and (3b) hold exactly for all workers so no additional rigidity is imposed. The point is that ensuring that appropriate incentive mechanisms are present does not imply that additional wage variation is introduced. Below, it will be shown that incentive wage schemes can actually reduce the amount of variation in an economy. First, it is useful to consider incentive schemes that employ the fact that workers are generally with the firm for more than one period.

IV. Life Cycle Incentive Devices

A number of authors have considered how age-earnings profiles can be altered in order to provide incentive effects. (See my 1979, 1981 articles; H. Lorne Carmichael, 1981; Carl Shapiro and Joseph Stiglitz, 1984; and Peter Kuhn, 1982.) The nature of the imperfect observability of output usually takes a somewhat different form here. Suppose that work takes place in two periods, but that output is not observed until the end of the period, and then only imperfectly, characterized for current purposes by occurring with probability p. All that is necessary, it turns out, is that one can observe that $u < u^*$ (or that $q < q^*$), the exact magnitude of the deviation being irrelevant.

One possible scheme that achieves a first best solution is to pay a wage W_0 in period zero and W_1 in period 1 if output in period zero was not observed to have fallen below q^*. If output in period zero is observed to have fallen below q^*, then the worker is terminated and is not permitted to work during period 1. For simplicity, assume that the discount rate is zero and that the alternative use of time in period 1 is zero. Then

$$(5)\quad W_1 = C(u^*)/p + x,$$
$$W_0 = Vu^* - W_1 = Vu^* - C(u^*)/p - x$$

will ensure that a first best equilibrium is attained for any $x > 0$.

To see this, note that the worker always works at zero effort in period 1 because there is no benefit from doing otherwise. His choice for period zero is either to work at zero effort or at effort equal to u^* because no intermediate value affects his income. He chooses to work at $u = u^*$ iff $W_1 - C(u^*) + W_0 > (1-p)W_1 + W_0$. This condition and zero profits imply equation (5).

What are the implications of such a scheme for wage rigidity? First, this scheme causes the age-earnings profile to deviate from the age-productivity profile. Output in period zero is Vu^*, but W_0 falls short of that. Output in period 1 is zero, and W_1 is necessarily positive. This does not imply, however, that wages vary more over the life cycle as a result of such a scheme. For example, if $Vu^* > 2C(u^*)/p$, then it is possible to find an $x > 0$ such that $W_0 = W_1$. If workers were paid their exact output in each period, the wage in period one would fall short of that in period zero. The incentive scheme would actually smooth earnings over the life cycle in an absolute sense.

How do the wages respond to business cycle fluctuations? The wage received over the entire life cycle is Vu^* so any permanent change in V is reflected one-to-one in the lifetime wage. But this does not imply the same correspondence in each of the two periods.

First, differentiation of W_0 with respect to V yields

$$\begin{aligned}\partial W_0/\partial V &= u^* + V\partial u^*/\partial V \\ &\quad - (C'(u^*)/p)(\partial u^*/\partial V) \\ &= u^* + V\partial u^*/\partial V - (V/p)(\partial u^*/\partial V),\end{aligned}$$

which is smaller than the right-hand side of (3a) because $\partial u^*/\partial V > 0$. The conclusion is that the wage in period zero does not move with the business cycle by as much as it would if output were geared directly to productivity.

For period 1, there is no change in productivity, but there is a change in the wage W_1. Differentiation with respect to V yields

$$\partial W_1/\partial V = (C'(u^*)/p)(\partial u^*/\partial V)$$
$$= (V/p)(\partial u^*/\partial V) > 0.$$

Although W_1 is more sensitive to changes in value of output than would be warranted by productivity considerations, it is still true that in neither period does the wage move as rapidly as the product price. This might give the appearance of wage rigidity since no one worker's wage at any point in the life cycle moves as rapidly as product price.

V. Relative Comparisons

More recent literature (see Sherwin Rosen and I, 1981; Jerry Green and Nancy Stokey, 1983; and Barry Nalebuff and Stiglitz, 1983) has discussed the role of relative comparisons in providing incentives. Tournament-type labor contracts, where one worker competes with another for a particular job that has a high wage, can induce workers to behave appropriately and to select first best levels of effort.

At this point, the variance of e in equation (1) can no longer be assumed to be zero. The essence of contest-like labor contracts requires that there be some random noise in the world. Without going into the details of the tournament labor contract, the basic idea is this: workers compete against one another for a particular job that carries a specified wage. The individual who has the highest level of output is given the job and is entitled to the wage that goes along with it, irrespective of the output level. It turns out that in competition, the equilibrium wage structure generates a first best solution, with each worker putting out the efficient amount of effort. The spread between the winner's wage and the loser's wage is the motivating factor.

If each individual draws an e in equation (1), the $e_i - e_j$ defined as z is distributed with density function $g(z)$. If W_1 is the wage that goes to the winner and W_2 is the wage that goes to the loser, then the equilibrium is[2]

$$W_1 = (V)(u^* + 1/g(0)), \tag{7a}$$

$$W_2 = (V)(u^* - 1/g(0)). \tag{7b}$$

First consider the wage variation relative to output variation. The expected wage across the two individuals is Vu^* and expected output from (1) is also Vu^*. However, no individual is paid the expected output. Individuals are *ex ante* identical, yet with certainty they receive different wages. In an *ex post* sense, output may have more or less variance across individuals than the wages. Since q is a random variable, whereas W_1 and W_2 are fixed in advance, whether actual q is more or less dispersed than wages depends upon the realization of the random variable z. Even the expectation of Vq_j, given that j is the winner, may be closer or further from Vu^* than is the winner's wage.[3] The conclusion is that the contest-type incentive structure adds variance in wages relative to the *ex ante* expected output, but may add or reduce variance relative to *ex post* output.

Additionally, the winner's wage is more sensitive, while the loser's wage is less sensitive, to changes in V than indicated by (3a). Differentiation of (7a) and (7b) yields

$$\partial W_1/\partial V = u^* + V\partial u^*/\partial V + 1/g(0),$$
$$\partial W_2/\partial V = u^* + V\partial u^*/\partial V - 1/g(0).$$

The last term on the right-hand side distinguishes these expressions from (3a), adding and subtracting wage flexibility, respectively. The reason is that when V increases, a

[2] See my article with Rosen.

[3] A simple example makes this clear. Suppose that e can take the values a and $-a$ only. Then $z = a_i - a_j$ has $g(0) = 1/2$ independent of the value of a so $W_1 = Vu^* + 2V$ and $W_2 = Vu^* - 2V$. What is the expected level of output, given that j has drawn $e_i \geq e_j$? It is $Vu^* + Va/2$ which can exceed or fall short of W_1, depending on the value of a.

higher u^* is appropriate. That can only be motivated by increasing the spread between W_1 and W_2. The first two terms on the right-hand side reflect the value of increased average productivity. The last term is the effect of increasing the spread.

"Winners" are the individuals who have advanced further up the hierarchy. This implies that salaries should be more volatile with output at higher job levels than at lower job levels. It is surely true that executives' compensation is more likely to be contingent on the performance of the firm than middle management's. Whether this relationship holds over the business cycle as well can be discovered.

VI. Conclusion

There exists no obvious link between wage rigidity and the provision of incentives. Some incentive devices introduce additional wage variation into an economy, but others, such as relative compensation schemes, can actually reduce it, even when risk aversion is not an issue. The more specific conclusions are:

Piece rates increase cross-sectional wage variation relative to a straight salary, but piece rate compensation is not necessarily more flexible over a business cycle than is a salary.

If output is not perfectly observed at zero cost, then some type of sampling scheme may be used. One reasonable scheme adds cross-sectional variation in wages. That same scheme makes some worker's wage rigid with respect to the business cycle, and others overly sensitive to the business cycle. But this is not a necessary consequence of imperfect observability. Another efficient scheme adds no variation that would not be present were piece rates and perfectly observed output to prevail.

Life cycle incentive devices steepen the age-earnings profile relative to the age-productivity profile. However, this does not imply that wages are more variable than productivity over the life cycle because the relative steepening could actually flatten the wage path. What is true, however, is that the annual wage is never as flexible with respect to the business cycle as are product prices.

Incentive schemes that involve relative comparisons, and in particular, tournament-style labor contracts, add wage variation relative to expected output. Such schemes do not unambiguously increase wage variation relative to realized output. An implication of this incentive scheme is that wages at the top of the hierarchy are more sensitive to changes in the value of output than wages at the bottom of the hierarchy.

Researchers who have conjectured that the provision of incentives can change the amount of wage flexibility in an economy are quite correct. Unfortunately, the direction of the change is not unambiguous. Those seeking to explain international or intertemporal differences in wage rigidity are not likely to find the answer in the incentive structure.

REFERENCES

Carmichael, H. Lorne, "Firm Specific Human Capital and Seniority Rules," Working Paper No. 101, National Bureau of Economic Research, February 1981.

Gordon, Robert, "Why U.S. Wage and Employment Behavior Differs from that in Britain and Japan," *Economic Journal*, March 1982, *92*, 13–44.

Green, Jerry R. and Stokey, Nancy L., "A Comparison of Tournaments and Contracts," *Journal of Political Economy*, June 1983, *91*, 349–64.

Hall, Robert and Lazear, Edward, "The Excess Sensitivity of Layoffs and Quits to Demand," *Journal of Labor Economics*, April 1984, *2*.

Holmstrom, Bengt, "Moral Hazard in Teams," *Bell Journal of Economics*, Autumn 1982, *13*, 324–40.

Kuhn, Peter, "Malfeasance in Long Term Employment Contracts: A New General Model with an Application to Unionism," Working Paper No. 1045, National Bureau of Economic Research, December 1982.

Lazear, Edward P., "Why is There Mandatory Retirement?" *Journal of Political Economy*, December 1979, *87*, 1261–64.

______, "Agency, Earnings Profiles, Productivity and Hours Restrictions," *American Economic Review*, September 1981, *71*,

606–20.

______ and **Rosen, Sherwin,** "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, October 1981, *89*, 841–64.

Nalebuff, Barry J. and Stiglitz, Joseph E., "Prizes and Incentives: Toward a General Theory of Compensation and Competition," *Bell Journal of Economics*, Spring 1983, *14*, 21–43.

Shapiro, Carl and Stiglitz, Joseph E., "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, June 1984, *74*.

Implicit Contracts, Explicit Contracts, and Wages

By ROBERT J. FLANAGAN*

Over the past decade, the search for an explanation of aggregate money wage stickiness in the face of substantial shifts in the marginal revenue product of labor—of the kind that was associated with an increase in real wages in the depression of the 1930's—has led to an extensive examination of explicit (union) and implicit (nonunion) employment contracting arrangements. This distinction can be exaggerated: long-term union contracts have an implicit element in the sense that they specify only a limited number of contingencies to which wages will be adjusted. Moreover, the implicit contracting branch of the literature has pushed in the direction of implying that the explicit contracts observed in the union sector may be formalizations of common informal arrangements in the nonunion sector, so that differences between the sectors in the behavior of wages and employment may be more apparent than real. The purpose of this paper is to examine the empirical basis for some of the key propositions of implicit contract theories and to note some important differences in the outcome of implicit and explicit contracting.

I. Contracting Approaches

The literature on implicit contract rests on the assumption that aggregate wage stickiness reflects wage inflexibility that is established in employment arrangements rather than aggregation phenomena, such as cyclical variations in the skill distribution of employment. Recent evidence on wage behavior at the establishment level confirms this assumption but indicates that the relative inflexibility of wages has developed since 1933 (see Daniel Mitchell, 1983).

The wage provisions of long-term collective bargaining agreements are one potential source of wage stickiness at the micro level. The duration of such agreements, the fact that they are rarely contingent on current influences on the economic fortunes of the firm, and wage interdependencies between unions and between the union and the nonunion sectors can insulate union wages from variations in the marginal product of labor. The difficulty is that the scope of the union sector is small in comparison to the presumed extent of wage rigidity. Union wages have accounted for perhaps 30 percent of the national wage bill in the past decade, and their importance was considerably smaller at the outset of the Great Depression. Moreover, the extent to which wages established in collective bargaining influence nonunion wage policies is questionable. Unlike the situation in many European countries, there is no legal extension of the terms of a collective bargaining agreement to nonunion firms, and empirical evidence indicates that union wages are more likely to be influenced by nonunion wages than the opposite (see my 1976 article).

Three general categories of explanation have emerged to explain why nonunion firms and workers have incentives to develop implicit contracts that produce wage rigidity. The literature on *risk shifting* (for example, see Martin Baily, 1974) assumes that employees are more risk averse than employers. A deal can thereby be reached under which an employer provides wage insurance to employees against changes in the economic "weather," but not changes in the "climate," a distinction that will be considered in more detail below. In general, this results in a gap between the wage and the marginal product reflecting the net insurance payment. In the *contracting* literature (for example, see Benjamin Klein et al. 1978; Michael Wachter and Oliver Williamson, 1978), the incentive to form contracts arises from the costs (particularly those associated with labor turnover) of an auction market for labor and the need for an agreement on how the surplus created

*Graduate School of Business, Stanford University, Stanford, CA 94305. Financial support was provided by the Stanford Graduate School of Business.

by specific training investments is to be split. Difficulties in specifying and monitoring all contingencies lead the parties to implicit understandings, while asymmetrical information pertaining to the reasons for proposed wage changes move the parties toward relatively inflexible wages as a protection against opportunistic behavior. Risk aversion is not required. In the *principal-agent* literature (for example, see Edward Lazear, 1981), the purpose of implicit contracts is to provide performance incentives to career employees. The gap between a wage and marginal product that arises in these cases reflects a performance incentive in a life cycle compensation arrangement in which an employee is induced to work harder (thereby raising the life cycle productivity profile) by receiving wages that first fall short of and then exceed current marginal product. Workers who accept such a contract are in effect offering a performance bond to an employer. Neither risk shifting, mobility costs, nor specific human capital are necessary for the principal-agent stories.

II. Wage Rigidity

Each of these categories of explanation offer a much richer understanding of incentives facing employees and employers in the nonunion sector than we had a decade ago. At the moment, however, there appears to be more empirical evidence on the preconditions for (for example, findings that typical employment relationships in the U.S. economy are lengthy; Robert Hall, 1982), than on the outcomes of implicit contracts. In particular, there are at least three difficulties in the implications of the literature for wage behavior. First, the literature rationalizes *real* rather than *money* wage rigidity (Costas Azariadis and Joseph Stiglitz, 1983).

Second, although the motivations differ, each of the implicit contracting stories reviewed above breaks the nexus between the real wage and the marginal product. In each of the stories, some *wage smoothing* will occur. But how much wage smoothing in fact occurs in modern economies? To what extent is the wage disconnected from the marginal product of labor? The answers to these questions are of crucial importance not only as tests of the major implication of implicit contracting theories, but also because they have fundamental implications for our views of the theoretical role of wages in allocating labor and, in empirical work, our interpretation of published wage statistics. If wage smoothing is extensive, so that wages primarily reflect long-term obligations of employers toward their employees, then the allocational role of wages is more limited than is implied in many models of labor market behavior. Similarly, if wage statistics are substantially influenced by "wage smoothing," how much can they tell us about current developments in the labor market? For a reading on current developments, it is the prices of labor being established in new contracts that become relevant—not the average prices for all operative contracts that characterizes most published wage data.

There has been surprisingly little examination of this basic question to date. The few exceptions to this statement find little support for extensive wage smoothing. James Brown (1983), for example, does not find evidence of significant wage growth following the completion of on-the-job training. Even if some wage smoothing occurs, it apparently falls considerably short of the smoothing provided by explicit (union) contracts. When wage adjustment equations are estimated for the union and nonunion sectors, a consistent finding is that the responsiveness of money wages to unemployment is significantly *higher* in the nonunion sector (see my earlier article).

A third difficulty concerns the extent to which employment arrangements offer the real wage insurance implied by the risk-shifting approach to implicit contracts. There is considerable evidence that employees simply do not receive this type of insurance generally and that employees under explicit contracts (who are presumably better able to obtain and enforce the kind of employment arrangement that they desire) do somewhat better than employees in the implicit contract sector. Formal cost-of-living adjustment provisions are hardly ubiquitous in either the union sector (where they are found in 50–60 percent of major private-sector collective bargaining agreements and 15–20 percent of public-sector agreements) or the nonunion sector (where they are found in about 4 percent of manufacturing firms). Even where they are found, it is clear that employers and employees share the risk of real wage variations attributable to unexpected price variations. COLAs rarely provide full compensation for changes in consumer prices. Wage-adjustment equations estimated separately for the union and nonunion sectors confirm that the extent of wage compensation for prices is even lower under implicit contracts than under explicit contracts. Employers are obviously unwilling to guarantee real wages over the length of the employment relationship as the risk-avoidance version of implicit contract theory implies. These findings simply reflect the tension between the empirical puzzle of rigid money wages and theoretical stories rationalizing rigid real wages.

III. Contract Enforcement and Termination

One of the more difficult issues facing the implicit contract literature has been the question of enforceability. Explicit labor agreements in the union sector are backed by law, which provides specific penalties to either party if they break a contact during its term. In addition the union serves as the agent of the workers to monitor compliance with the contract and, through the grievance procedure, can protest and seek correction of apparent deviations from contract terms (as can management). While disagreements over the interpretation of a contract are frequent, abrogation of a contract is rare. Employers in the nonunion sector are under no such constraints. Under the traditional common law, nonunion employees serve "at the will" of their employers, with the result that until very recently, there have been few legal barriers to breach of implicit labor agreements by employers or workers. How and why are implicit contracts broken? If implicit employment arrangements mimic explicit contracts, is the pattern by which contracts are terminated the same in the two sectors?

Under the principal-agent analysis, relatively steep age-earnings profiles provide stronger incentives to work and a greater "bonding" of workers to the firm. At the same time, a steep age-earnings profile creates incentives for firms to renege on the contract and lay workers off before they have collected their wage premia in the later part of their life cycle. Therefore, the steeper the age-earnings profile, the higher the probability of contract termination via layoff and the lower the probability of contract termination via a quit. Conversely, relatively flat age-earnings profiles are more likely to be associated with quits than with layoffs.

It is fairly well established that age-earnings profiles are steeper in the nonunion (implicit-contract) sector than in the union sector. Moreover, the pattern by which contracts are terminated differs between the two sectors. Nonunion workers are more likely to quit and less likely to be laid off their jobs than are union workers (Richard Freeman, 1980; James Medoff, 1979). That is, workers are more likely to break implicit contracts, while employers are more likely to break explicit contracts. Comparative turnover behavior (after controlling for general influences) appears to be just the opposite of the predictions of the principal-agent approach. Given a union-nonunion wage differential, however, the pattern of terminating contracts in the two sectors is consistent with the contracting branch of the literature. Employees are less likely to leave jobs that reward them well relative to their alternatives on average (because they collect more of the surplus created by specific human capital). Employers are more willing to use layoffs when they collect a relatively small fraction of the surplus. The pattern of layoffs between the two sectors also reflects the differences in the degree of wage rigidity noted in the previous section.

Thus far, implicit contract theories have relatively little to say about why different employment arrangements might emerge under explicit and implicit contracts. Indeed, the thrust of the literature has been in the other direction. Recent thinking about explicit labor agreements, however, has stressed that with democratic procedures for contract ratification and the election of union leaders, union policy is essentially determined by the preferences of the median union voter (rather than the marginal worker in an auction labor market). Moreover, by allocating employment opportunities (including protection from layoffs) on the basis of seniority, unions can establish employment contracts in which a policy of wage rigidity is de facto a policy of *income* security for the median voter under normal cyclical events. Therefore, the median voter can support a policy of wage inflexibility knowing that seniority offers protection against layoffs and employment variability over normal cycles (changes in the weather).

IV. Adjusting the Changes in Climate

All of the general approaches to implicit contracting imply wage smoothing through changes in the economic weather, but the models are generally vague on their implications for adjustments of contractual relationships in response to changes in the economic climate.

The contracting literature generally motivates wage rigidity as a defense against opportunistic behavior when specific skills create a bargaining surplus and when the costs of monitoring and continuous performance appraisal are high. Efforts to change the wage unilaterally on the basis of developments that were not public information would create suspicion. (The principal-agent literature with asymmetric information would recognize a similar problem.) A rigid wage policy mitigates suspicion of opportunistic behavior in (typical) situations where parties are unable to specify all contingencies in formal contracts.

When changes in the economic climate, such as occurred in the earlier 1980's, increase the wage variability that would be forthcoming in a spot market for labor, the cost of a system of relatively inflexible wages increases for both parties. But simply dropping rigid wage (implicit) contracts also makes it easier for a firm to cheat by opportunistically lowering wages by more than shifts in demand would warrant. The contracting literature emphasizes vertical integration as a solution to this problem (see Klein et al.). In the context of the labor market problem this implies an expansion of employee-ownership institutions. When the worker is the owner, the agent is the principal, and the incentives for opportunistic behavior diminish. Therefore, one implication that might be read into the new literature is that stark changes in climate would appear to build pressures for employee-ownership arrangements.

Interestingly, this same conclusion follows (for different reasons) for the basic model of union decision making over explicit contract objectives. Only when there is an abnormally deep cycle so that layoffs reach far enough down the seniority list to threaten the employment of the median voter does wage flexibility (concessions) or an employee-ownership arrangement become feasible. In less severe times, when seniority protects the median union member from layoff, general wage adjustments or worksharing (hours variability—an adjustment that is more prevalent in the nonunion sector) are opposed because they threaten the income security of the median member.

V. Conclusions

Over the past decade, the literature on implicit labor agreements has provided a number of illustrations of how privately efficient employment contracts might generate wage rigidity. Empirically, however, several key features of implicit contract theories appear to be open to question. First, while there is strong evidence of long-term employment relationships between employers and employees, we do not know the extent to which these relationships result in wage smoothing. There is evidence, however, that wage smoothing is greater under employment arrangements established through collective bargaining. Second, the literature provides theoretical support for rigid real wages, whereas rigid money wages is the phenomenon to be explained. In practice, neither implicit nor explicit contracts typically provide complete real wage insurance, and to the extent it is provided, union members receive more compensation for price changes than nonunion members. Third, the way in which the employment relationship is terminated differs under explicit and implicit contracts. Fourth, prevailing theories of both explicit and implicit contracting predict increased interest in employee-ownership institutions in response to a change in economic climate. Finally, there is some evidence that nonunion money-wage rigidity at the establishment level is a phenomenon of the postwar period. This poses an interesting research question of how theories of implicit and explicit contracts account for changes in the degree of wage flexibility over time.

REFERENCES

Azariadis, Costas and Stiglitz, Joseph E., "Implicit Contracts and Fixed Price Equilibria, *Quarterly Journal of Economics*, May 1983, *98*, 1–22.

Baily, Martin N., "Wages and Unemployment Under Uncertain Demand," *Review of Economic Studies*, January 1974, *41*, 37–50.

Brown, James N., "Are Those Paid More Really No More Productive?," Working Paper No. 169, Industrial Relations Section, Princeton University, October 1983.

Flanagan, Robert J., "Wage Interdependence in Unionized Labor Markets," *Brookings Papers on Economic Activity*, 3:1976, 635–73.

Freeman, Richard B., "The Exit-Voice Tradeoff in the Labor Market: Unionism, Job Tenure, Quits, and Separations," *Quarterly Journal of Economics*, June 1980, *94*, 643–73.

Hall, Robert E., "The Importance of Lifetime Work in the U.S. Economy," *American Economic Review*, September 1982, *72*, 716–24.

Klein, Benjamin, Crawford, Robert G. and Alchian, Armen, "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," *Journal of Law and Economics*, October 1978, *21*, 297–326.

Lazear, Edward P., "Agency, Earnings Profiles, Productivity, and Hours Restrictions," *American Economic Review*, September 1981, *71*, 606–20.

Medoff, James L., "Layoffs and Alternatives Under Trade Unions in U.S. Manufacturing," *American Economic Review*, June 1979, *69*, 380–95.

Mitchell, Daniel J. B., "Wage Flexibility: Then and Now," Working Paper No. 65, Institute of Industrial Relations, University of California-Los Angeles, December 1983.

Wachter, Michael L. and Williamson, Oliver E., "Obligational Markets and the Mechanics of Inflation," *Bell Journal of Economics*, Autumn 1978, *9*, 549–71.

RECENT DEVELOPMENTS IN THE THEORY OF INDUSTRIAL ORGANIZATION: SOME MACROECONOMIC IMPLICATIONS

Price Rigidities and Market Structure

By JOSEPH E. STIGLITZ*

Conventional wisdom has it that a large part of the explanation of Keynesian unemployment is the observed rigidities of wages and prices. What has been lacking, however, is a satisfactory theory (or conjunction of theories) which explains how wages and prices can be at non-market-clearing levels.

The objective of this paper is to sketch out several alternative theoretical explanations of this phenomenon. Each (in contrast with Keynesian theory) is consistent with the observation that as the economy goes into a recession, real product wages do not increase, and may, in fact, decrease. All of the theories assume rational, profit-maximizing firms, and all entail some important departure from the standard competitive paradigm. The first three explanations represent minor adaptations of standard models.

I. Some Simple Explanations of Price Stickiness

A. *Long-Term Contracts*

The first explanation denies, in effect, the relevance of the observation. Firms have long-term relationships with their customers that entail (implicit or explicit) insurance; the firm charges less than the marginal cost of production in good states and more in bad. The buyer is assumed to be (in effect) less risk averse than the seller. Though it is undoubtedly important to take into account long-term contracts in the analysis of short-run movements in observed prices, I doubt whether these can fully account for the phenomenon I am attempting to explain. Why should buyers, on the whole, be willing to supply this insurance to sellers? Because workers, too, are on long-term contracts, spot wages and long-term wages differ; the theory predicts that spot real product wages increase during the recession. If long-term contracts are relatively more important in labor markets than in product markets, then the anomaly which is to be explained is even larger than suggested by data that fail to distinguish between spot and contract wages and prices.

B. *Learning by Doing: Declining Marginal Cost Curves*

The second explanation denies the hypothesis of diminishing marginal returns. I have identified one (and only one) important instance in which the *marginal* cost is reduced as the level of production increases, and that occurs when there is learning by doing.

This explanation may be important in a few sectors, but the phenomenon that is to be explained, the failure of real product wages to rise by very much, is hardly limited to sectors where learning is important.

C. *Increasing the Elasticity of Demand*

If there is diminishing marginal returns and the economy is competitive, the real product wage should rise as the economy goes into a recession. The same is true if the economy is dominated by monopolies (or monopolistically competitive firms), who set marginal revenue equal to marginal cost; so long as the elasticity of demand remains unchanged, price will remain a given markup

*Department of Economics, Princeton University, Princeton, NJ 08544. Financial support of the National Science Foundation is gratefully acknowledged. I am indebted to Andrew Weiss, Bruce Greenwald, Franklin Allen, Robert Levin, Richard Gilbert, and Steven Salop for many discussions on the subject of this essay. Space limitations prohibit providing compete references for the relevant literature.

over marginal cost. If, however, the elasticity of demand were to decrease, then the markup might increase; and thus, even if marginal costs of production fell, prices would not fall proportionately, and might even rise. The question is, is there any reason to believe that in a recession the elasticity of demand is decreased? Though there is no general presumption for a decrease in the elasticity of demand, two cases warrant mentioning. First, consider a monopolistically competitive market, say, characterized by n firms distributed around a circle. Assume that initially, demand is sufficiently strong that all market areas are served, and the equilibrium price thus is affected both by the intensive margin (the response of current customers to changes in price) and the extensive margin (the additional customers who purchase at the given store as a result of a price reduction). If the demand curves of each individual shifts enough, equilibrium may be characterized by some areas not being served by any store (i.e., the price, including transportation costs, is sufficiently high that their demand, with the new demand curves, is zero). Since there is then only an intensive margin, the equilibrium markup over marginal costs will now be higher.

Assume now that search is costly. Consider the product variety interpretation of the standard circular monopolistically competitive model. Individuals have a preferred point (product variety) on the circle; the utility they get from a product decreases the greater the distance of the product from the preferred point. Assume individuals randomly arrive at points along the circle, and there is a fixed cost for each additional search. As usual, they will have a simple reservation-quality-price rule: if $v(p,q)$ gives their level of utility as a function of price and quality, then they purchase at a store if $v(p,q)\geq \hat{v}$, for some critical $\hat{v}$. The critical value of $\hat{v}$ depends on a number of factors, including the market rate of interest and the cost of search. If the real rate of interest rises in a recession, v will decrease; each individual will find a wider range of products acceptable. This will *reduce* the elasticity of demand, since there will then be fewer individuals who are just at the margin between buying at one store and continuing to search; and this will increase the price (relative to marginal cost) charged at each store.

The discussion of the preceding three sections has shown that modifications of the standard theory may provide some insights into wage stickiness, but hardly provide a convincing general explanation. In the next sections, I turn to some alternative explanations based on more fundamental departures from the standard paradigm. The first two are based on imperfect information in a competitive environment, the second two on strategic behavior in a noncompetitive environment.

II. Imperfect Information and Price Stickiness

A. *Judging Quality by Price*

The first of the imperfect information theories is based on the premise that customers are imperfectly informed about the characteristics of the products which they purchase, at the time they purchase them. There is a widespread belief that higher priced commodities are of higher quality. This may be either because of selection effects (for example, potential sellers of good used cars will not sell them unless they are offered a high enough price), or incentive effects (if price exceeds marginal cost by enough, it pays to maintain one's reputation by producing high quality commodities). Lowering one's price may be interpreted as a lowering of quality, and thus demand may actually *decrease* rather than increase. In that case, firms may not respond to situations where the value of the marginal product of labor exceeds the wage by lowering price, as traditional theory would have predicted: it *appears* as if the firm is off its supply curve.

A simple example of a reputation model may help illustrate this point. A firm sells a product at a price p; the marginal cost for producing the high quality product is c_h for the low quality c_l. Individuals cannot assess the quality of the product until after they have purchased it; they continue to purchase from the given firm unless he cheats on them, and sells them a low quality commodity. Once they have been cheated upon, they refuse to purchase from the given firm again. The present discounted value of profits of the firm, if it remains honest, is $p - c_h/r$, where r is the rate of interest; the (present discounted value of) profits of the firm if it cheats is $p - c_l$, it gets its sales this period, but nothing thereafter. For the firm not to cheat,

$$(1) \qquad (p - c_h)/r \geq p - c_l$$

$$\text{or} \qquad p \geq (c_h - rc_l)/(1 - r).$$

Thus, even in a competitive economy, price will exceed the marginal cost of production. If the *real* interest rate increases as the economy goes into a recession, then the markup over marginal costs will increase.

Two modifications of this argument strengthen the presumption that prices may not fall (relative to wages) in a recession. First, assume that firms die exponentially, at the rate μ. Second, assume that the firm treats all customers the same (it cheats one, it cheats them all; this would be optimal if reputations spread quickly across customers). Let $Q = F(L)$ be the firm's output. Then the condition for producing high quality commodities becomes

$$(2) \qquad p \geq (LW/Q)/(1 - (\mu + r)),$$

where W is the wage. Since death rates among firms are higher in a recession, it is reasonable that consumers will perceive μ as higher in such a situation, and thus price must increase to ensure that the firm produces a high quality commodity. Second, (2) makes it clear that what is relevant is the markup over average (variable) costs; if there are fixed costs to a firm, it is possible that there will be diminishing marginal products and increasing average products (this is consistent, for instance, with Okun's Law); in that case since the average productivity has decreased, the equilibrium price (relative to the wage) will need to increase to maintain the firms' incentives to produce high quality commodities.[1]

The phenomenon I have just described is much more general than the specific reputational model I have discussed. A similar result obtains if current prices affect not only present but future demand. Firms trade off lower prices today for higher future sales; an increase in the cost of capital associated with a recession implies that the increased future profits will be less valuable, and hence firms will increase their current markup over marginal costs. (See my paper with Bruce Greenwald and Andrew Weiss, 1984).

B. *Asymmetric Responses with Costly Search*

It has long been recognized that if the demand curve facing a firm is kinked, then the firm may not respond to changes in the factor prices it faces. Thus, if it is hypothesized that the demand curve shifts to the left by z percent, with the kink remaining at the same price, then output will be decreased by z percent, but price will remain unchanged. If the kink occurs at a higher price, then price will increase; we would then observe the seemingly paradoxical result of increasing prices and falling output. Though some recent work in game theory has attempted to put the old oligopoly kinked demand curve on more solid footing, we are concerned here with kinked demand curves which arise in competitive markets because of costly search. Consider a market equilibrium in which all firms charge the same price. If a firm raises its price, all of its customers know that it has raised its price, and those with low search costs may proceed to search for one of the lower priced stores. If a firm lowers its price, it may sell more to its current customers; and more individuals who are searching who happen to stop at the store will decide to purchase there. But even if it became pub-

[1] This paragraph is intended to present a simple heuristic explanation of why reputation models can generate observed patterns of movements in real product wages. It is easy to show that the reputation equilibrium as described can arise as a perfect equilibrium in a repeated game. The formalization embodied in equations (1) and (2) assumes that current value variables will persist indefinitely; it is easy to extend the analysis to assume that there are, say, two states of nature ("good" and "bad") with stationary transition probabilities between the two. See Franklin Allen (1983). If consumers are imperfectly informed concerning current variable costs, and are risk averse, the critical price may be completely rigid.

licly known that *some* store had lowered its price, it might not become known which store had lowered its price; if there were many stores in the market, those at other stores, even with relatively low search costs, might thus not be induced to try to find the store. Thus, the store may gain fewer customers when it lowers its price than it loses when it raises it price, giving rise to a kink. The kink may well change as the environment changes; if, as the economy goes into a recession, all consumers believe that other stores have lowered their price by v percent, then the kink will occur at a price which is lower by a corresponding amount; for if the firm does not lower its price by that amount, the low search cost customers will be induced to search. At the same time, if they believe that all other firms have raised their price by z percent, the kink will occur at a price which is higher by z percent. Only if the firm increases its price by more than z percent will they be induced to search. There is a fundamental indeterminancy in the location of the kink (and hence in the equilibrium price); there are a multitude of rational expectations equilibria, including some which exhibit stickiness of money wages. (For a more extensive development of these ideas, see my 1983 paper.)

III. Oligopolistic Theories

Because of the plethora of possible patterns of interactions of firms in oligopolistic markets, economists have been loath to use oligopoly theory to provide insights into macroeconomic behavior. This seems to be excessively cautious: we do not need to claim that a particular oligopoly model describes behavior in all oligopolistic industries; only that it provides insights into the behavior of some. In this section, I wish to suggest how two recent developments in oligopoly theory can provide some insights into price rigidities.

A. *Limit Pricing and Entry Deterrence*

There is a long tradition that has held that monopolists kept lower prices than they otherwise would in order to deter entry. Potential competition may be (almost) as effective as actual competition in keeping prices down. As the economy goes into a recession, the extent of excess capacity increases. Excess capacity is an effective deterrent. Because of the availability of excessive capacity, firms do not need to rely as much on limit pricing to deter entry.

This argument, though plausible, has two critical flaws: neither prices nor excess capacity may serve as an entry deterrent. For low prices to serve as an entry deterrent, or for high prices to encourage entry, implies that potential entrants believe that the incumbent firm(s) will leave their prices unchanged in response to entry (or at least that the price they charge after entry will be correlated with their pre-entry price). As stated, the analysis does not provide an explanation for why this should be so. Secondly, as Dixit has emphasized, for excess capacity to serve as an entry deterrent requires that entrants believe that it will be used after entry. This will not be the case if the post-entry equilibrium is a Nash-quantity equilibrium.

Still, one suspects, there is something to the argument that in a recession, the threat of entry is less important, and this allows firms to charge higher prices (relative to marginal costs). The following model (an adaptation of one due to Salop) provides one context in which this can be shown to be true.

Assume the potential entrants do not know what the marginal cost function of the incumbent firm is; they know that it may either be c_1, or $c_2 > c_1$. The potential entrant believes it will be profitable to enter if his rivals costs are c_2, but not otherwise. The incumbent, knowing this, attempts to convince the potential entrants that his costs are low. The potential entrants, of course, know this. We have a standard screening self-selection problem. The equilibrium may easily be characterized. If the incumbent has a high cost, he charges a price, p_2^* such that marginal revenue equals marginal cost. If the incumbent has a low cost, he charges a much lower price: sufficiently low that the high-cost firm does not try to imitate him. Let $\pi(p, c)$ denote profits as a function of price charged and marginal cost of production. We simplify the analysis by assuming only two periods with $\delta = 1/(1+r)$ being the discount factor. Assume that the probability of entry the second period (if the potential entrants know that the incumbents costs are c_2) is λ, and that profits of the incumbent upon entry are (for simplicity) zero. Then the equilibrium prices charged by the low-cost store satisfies

$$(3) \quad \pi(p_2^*, c_2)(1+\delta(1-\lambda)) = \pi(p_1, c_2)(1+\delta).$$

It immediately follows that a reduction in the flow of entry (λ) or an increase in the rate of interest (a decrease in δ), will result in an increase in p_1. (An increase in the threat of entry makes the high-cost firms more willing to sacrifice current profits, to imitate the low-cost firms.) Thus, once again, a recession may be associated with an increase in markup over marginal costs.

B. *Coordinating Collusive Behavior*

Recent developments in the theory of repeated games have shown how collusive behavior may be sustained (as perfect equilibrium) in noncooperative games. Individuals who are detected to deviated from cooperative behavior are punished.[2] It is often difficult, however, to detect deviations from cooperative behavior, particularly when demand curves are random and differ from firm to firm. Consider a cartel in which the price charged by each firm is observable,[3] but that is all. One set of (perfect) equilibrium strategies entails each firm charging the cooperative price, p^*, so long as the rival charges p^*; if any firm deviates, however, all firms revert to the Nash-equilibrium prices, p^N, for an extended period, after which prices again return to the cooperative level.[4] The cooperative price p^* is chosen to maximize the expected (present discounted value) of profits, given the demand variability (and given that the price cannot be changed in response to the change in demand). In this perfect equilibrium, firms do not respond to changes in the environment. Each firm cannot detect the changes in other firms' demand curves. They cannot discriminate between cases when rivals lower their price to cheat on the collusive arrangements, and when rivals lower their price in response to a reduction in their demand. The gains from being able to enforce collusive behavior outweigh the losses from failing to adapt the price to shifts in the demand curve.

IV. Concluding Remarks

This paper began by observing that the following statements are not consistent: (*i*) firms are competitive and on their supply curve; (*ii*) real product wages did not increase when employment decreased (technology and capital remaining constant); and (*iii*) there are diminishing returns to labor.

One of these hypotheses must be false. This paper is based on the premise that the empirical observations, the failure of real product wages to rise, is valid, at least for some important industries. I have provided one explanation for why the marginal product of labor might have decreased in the recession (learning by doing). If the sector is characterized by a technology exhibiting increasing returns, it will not be perfectly competitive. Thus, I have focused attention here on the first premise. If the market were noncompetitive, then real product wages could fall as the economy went into a recession, if the elasticity of demand decreased; again, I have provided two models (based on costly search and monopolistic competition) that generate that result.

Most of the explanations of the seeming paradox which I have offered reflect a more fundamental alteration in traditional theory: (*i*) if prices convey information, firms may be reluctant to lower prices lest consumers believe that there is a reduction in the quality of the product; in these circumstances, I have shown that markups over marginal costs may well rise in a recession; (*ii*) when search is costly, demand curves may have a kink in

[2] Whether effective punishments can be designed depends, of course, on the rate of interest. It is possible that if as the economy goes into a recession, the real rate of interest increases, it may no longer be possible to enforce collusive behavior, in which case we would find such markets behaving more competitively.

[3] In many important cases, this may not be true: posted prices and traded prices may differ markedly.

[4] Provided the discount rate is low enough and demand variability small enough, there will exist a perfect equilibrium of this form.

them; (*iii*) in oligopolistic markets, the reduction of the threat of entry in a recession may lead to an increase in price; and (*iv*) maintenance of collusive arrangements by noncooperative equilibria may entail price rigidities.

I suspect that some of these theories may provide a better description of some markets, others of other markets; there is no compelling reason to believe a single theory provides the explanation of price rigidities in all markets.

In these theories, firms all behave "rationally"; an alternative "explanation" of seeming price rigidities is that firms' managers act according to certain rules of thumb, for example, those that entail a markup over average costs. Though such "theories"—if they can be called that—fail to explain how or when the markups change, as they undoubtedly do, they may provide as good a description of the short-run behavior of the firm as our more sophisticated theories.[5]

Most of the theories I have presented have one thing in common: they are inconsistent with the competitive theory of supply which underlies the two strands of thought which have dominated research in macroeconomics in recent years (the rational expectations and fixed price models). Thus, whether predictions based on those models (for example, concerning consequences of policy changes) have any validity remains a moot question. What is needed is a macroeconomic theory based on theories of imperfect competition and imperfect information. This paper is intended as a contribution to the construction of that alternative paradigm.

[5] Other short-run dynamics may provide an explanation of the observed phenomena: firms do lower prices, but as quickly as they lower prices, wages fall, so that real product wages fail to rise. Observed movements on real product wages are thus attributable to assumptions concerning relative adjustment speeds; these explanations are unsatisfactory, since the assumptions concerning relative adjustment speeds are, to say the least, *ad hoc*.

REFERENCES

Allen, Franklin, "A Theory of Price Rigidities When Quality is Unobservable," mimeo., University of Pennsylvania, December 1983.

Stiglitz, J. E., "Duopolies Are More Competitive Than Atomistic Markets," mimeo., Princeton University, October 1983.

Greenwald, Bruce, Stiglitz, Joseph E. and Weiss, Andrew, "Informational Imperfections in the Capital Market and Macroeconomic Fluctuations," *American Economic Review Proceedings*, May 1984, *74*, 194–99.

Industrial Organization and Competitive Advantage in Multinational Industries

By A. MICHAEL SPENCE*

My purpose in this paper is to outline some of the developments in the study of industrial organization and evolution, and of competitive strategy that are pertinent to the evaluation or assessment of policies whose purpose is to change the pace or direction of industrial development, or the relative competitive position of "domestic" firms in a multinational industry.

Industries (even concentrated manufacturing industries) are sharply distinguished from each other by their structural characteristics; the latter affect competitive conduct, indeed, even what are the important dimensions of competition—evolution and dynamic performance. That proposition appears at face value to be rather obvious, but it is worth noting that it is inconsistent with studying "the oligopoly problem," or with basing policy on a sample division of industries into two categories, such as sunrise and sunset.

It is appropriate to ask what accounts for the differences in the relative performance of countries (or firms based in separate countries) in multinational industries. This question has achieved some immediacy because of the increased market share in the United States, and in worldwide markets achieved by non-U.S. multinationals. Generally, the rate of growth of trade flows (exports and imports) in many industries exceeds (often by factors of 2 or 3) the rates of growth of domestic output in *any* single country.

One of the lessons of the study of multinational industries is that there are numerous policies that influence them. They include (in no particular order): 1) policies that operate on domestic industry structure (horizontal concentration, vertical integration, specialization), 2) export promotion *and* designation of authorized exporters, 3) import restrictions, 4) restrictions on foreign direct investment flowing in and out of a country, 5) policies affecting interfirm transfers of technology within the domestic industry and multinationally, including licensing of technology, and the uses and protection of intellectual property, 6) policies that operate on costs or input prices, including interest rates, costs of capital, risk-spreading, investment subsidies (including those to *R&D*), 7) various other features of the tax system, and 8) policies designed to achieve some level of coordination or convergence of expectations among competitors within an industry and with government. The impact of these policy options can be observed in many detailed case studies.

I. The Market Failure/Competitive Advantage Approach

There are three categories of public sector activity that can be used by governments under certain structural conditions to increase their country's net surplus, profits, and share.

A. *The Strategic Use of Blocked Access to Domestic Markets*

The first category is blocking access to markets in industries with declining average costs. Scale economies of static and dynamic kinds characterize many components of costs. Their importance in a multinational industry is related to three structural features of the industry: the fraction of total costs accounted for by each component; the elasticity of unit costs with respect to volume (or in the case

*Department of Economics, Harvard University, Cambridge, MA 02138. This research was supported by the National Science Foundation, the Kennedy School of Government, and the Harvard Business School. The paper is a shortened version of "Industrial Organization, Competitive Advantage in Multinational Industries and Industrial Development Policies," HIER discussion paper, January 1984, which cites the related literature.

of the learning curve, accumulated volume); and whether the scale economies get truncated at national boundaries as tends to be the case in marketing and distribution, but not in *R&D*, and only partially in manufacturing. With respect to those components of cost in which there are substantially negative elasticities of unit cost with respect to volume, access to major markets or market segments is a significant part of establishing a firm's relative cost and competitive position. For example, if the elasticity of unit costs with respect to accumulated volume is 0.32, then costs go down by 20 percent each time volume doubles. Suppose there are three market areas, *A*, *B*, and *C*, representing 40, 30, and 30 percent, respectively. A firm based in *A* has the share S_1 in markets *A* and *B* (it is excluded from *C*). Firm 2 has the share S_2 of *A*, *B*, and *C*. The ratio of the unit costs of firm 2 to firm 1 is $(.7S_1/S_2)^{.32}$. The unit costs are the same if $S_2 = 1.43S_1$. Alternatively, if $S_1 = S_2 = 25$ percent, then firm 2 will have an 11 percent cost disadvantage, a matter of some competitive significance.

The *R&D* is a more complex subject than the sample analysis of scale economies might suggest, but scale economies are a part of the competitive relevance of *R&D*. The costs of achieving a given rate or level of product development or cost reduction are largely fixed. As a result, the elasticity of unit costs with respect to volume is minus one. Because *R&D*-induced scale economies tend not to be truncated by national market boundaries, share and market access is particularly important in *R&D*-intensive multinational industries. If, as is true in the technologically advanced sectors, *R&D* runs in the range of 10 to 15 percent of sales, then a relative market share of 0.5 may give rise to a 10 percent cost disadvantage by itself.[1] While the share of costs is relatively small, the elasticity of unit cost with respect to volume tends to be large (because the costs themselves are insensitive to volume). Thus in the *R&D*-intensive industries, there is a premium placed on having access to major markets, to obtain sales against which to amortize the *R&D* costs.

Blocking of access to a market or submarket may be needed to acquire a competitive cost position. Once a competitive cost position is achieved, blocked access will not usually be required to maintain a competitive cost position. Thus the access-blocking tactic should be seen in a strategic context as a device for lowering the cost of entry or expansion. The internal rate of return on the entry investment is *raised* by the blocking of access to the domestic market. Note that in industries with significant scale economies, entry is generally responded to vigorously by established firms because of the importance to them of *not* losing share. That response is largely responsible for making the entry costs high. Blocking access blunts or eliminates the response. Further, the net surplus (in present value terms) to the country that does it, *can be* positive.[2]

I should also add that blocking access to a domestic market is not always undertaken for the strategic reasons outlined above. It can be, and is, done simply to protect the domestic industry, which, absent the protection, would not survive in the long run. This last form of access control has a positive cost under most conditions for the country that undertakes it. That is to say, the consumers pay a price for the absence of foreign competitors in the domestic market.

It is also true that a policy of blocking access to reduce the entry cost for domestic competitors to allow them to develop a competitive cost position can fail to increase the domestic surplus. It will fail if scale economies are limited. It will also be ineffective if the developing domestic industry exploits the protection, not to expand share to reduce costs, but rather to appropriate the rents (in the form of increased costs) created by the protection from competition. And, finally, if access to foreign markets is blocked as a countermeasure, then the policy may result

[1] The relative market share of a firm is the ratio of its share to that of its largest competitor.

[2] Whether the net surplus is positive depends in part on whether, and by how much, blocking access to the domestic market raises prices to domestic consumers for some period of time.

in the acquisition of a competitive cost position, but not the capacity to export.

B. *The Use of Subsidies to Shift the Equilibrium in an Imperfectly Competitive Market*

It is well known that the relative costs of competitors influence their market shares and profitability. It is also true, that if the costs of a subgroup of competitors in an *imperfectly* competitive (i.e., oligopolistic) industry are subsidized, and if there are no countervailing subsidies provided to competitors outside the group, then the profits of the subsidized group *may* increase by more than the gross amount of the subsidy. The increase in profits results not only from the margin increase, but also the increase in market share.

There is then, under certain structural conditions, a potential *net* benefit to be obtained by lowering the costs of domestic competitors in a multinational industry via the subsidization of costs or input prices, but there are some important qualifications.

Suppose there are two countries, 1 and 2. There are n_1 firms in country 1 and marginal cost of c_1 (not volume dependent). I will assume here that the elasticity of demand in each country is β. The price in country 1 is p and in country 2 the price is q. The market in country 2 is ϕ times that in country 1. Here ϕ could be any number greater than zero. Country 2 firms have unit costs of c_2. It has two kinds of firms: one group is authorized to export, there are m_1 of them; the second group serves the domestic market and there are m_2 of them. The exporting firms receive a subsidy of $(1-\theta)$ on their costs for exports. All firms receive a subsidy of $(1-\delta)$ on domestic sales in country 2. Country 1 does not engage in subsidies, or export restrictions on particular firms, that is, it is passive. The equilibrium in each market is a Nash equilibrium in quantities.

For this analysis, I will focus on the net surplus in country 2 (including the consumer's surplus, the profits on domestic and foreign sales, and the subsidy, counted negatively). The equilibrium in this model is easy to calculate, and the calculations are of no interest. I proceed directly to the results.

The prices in country 1 are

$$(1)\quad p=(n_1c_1+m_1\theta c_2)/(n_1+m_1-1/\beta),$$

and, in country 2,

$$(2)\quad q=(n_1c_1+(m_1+m_2)\,\delta c_2)\,/\,(n_1+m_1+m_2-1/\beta).$$

The *net surplus* for country 2 is

$$(3)\quad T_2=\big[\phi/(\beta-1)q^{1-\beta}+\phi(m_1+m_2)\beta q^{-(1+\beta)}(q-\delta c_2)(q-c_2)\big]+\big[m_1\beta p^{-(1+\beta)}(p-\theta c_2)(p-c_2)\big].$$

The two terms in square brackets are the surplus in the domestic market (term 1) and the profits net of subsidies on export sales in country 1 (term 2).

The first thing to note is that these are separable in the sense that m_1 and θ can be set to maximize foreign profits without influencing term 1 at all. In particular, country size, ϕ, is of no relevance with respect to earnings on nondomestic sales. Note that this would not be true if we forced $\theta=\delta$, *or* if we added economies of scale in the form of marginal costs that decline as volume increases.

The second observation is that the maximum of term 2 (foreign earnings net of subsidies) can be achieved with $\theta=0$; that is, no subsidy. From the equilibrium price, we find that the earnings of exporting firms in country 2, per unit sold per firm, are

$$(4)\quad m_1(p-\theta c_2)=n_1c-(n_1-1/\beta)p.$$

Thus, upon substitution, term 2 in (3) becomes

$$(5)\quad E_2=\beta p^{-(1+\beta)}(p-c_2)\times(n_1c_1-(n_1-1/\beta)p).$$

That is, net earnings depend only on the price in country 2. That price will be affected by both m_1 and θ, but it doesn't matter which combination is chosen. In particular, country 2 can get all the benefits obtainable by setting the subsidy θ equal to zero, and then selecting the number of authorized exporters appropriately. Too many will drive the price in the foreign market down too far, and too few will exploit the profit potential too little.

The purpose of the model is not to dismiss subsidies. Surely if economies of scale were reintroduced, there might be a preference for subsidies over a proliferation of competitors, but competition is to some extent a substitute for subsidies in exploiting the benefits of foreign markets.

The actual optimum is of some interest. Absent competition from country 2, the price in country 1 would be

$$\bar{p} = n_1 c_1/(n_1 - 1/\beta). \tag{6}$$

If $c_2 \geq \bar{p}$, then country 2 can't compete profitably in country 1 and $m_1 = 0$. Otherwise, $c_2 < n_1 c_1/(n_1 - 1/\beta)$. In that case, profits rise and then fall to zero at $p = c_2$ as the price p declines starting at $p = n_1 c_1/(n - 1/\beta)$. The price declines monotonically as m_1 rises. It is easy to establish that $E_{pc_2} > 0$ so that the optimal price is an increasing function of c_2. Whether c_2 is greater than or less than c_1 is not directly relevant, except that for given c_1, a value of $c_2 < c_1$ will result in more competitors m_1 and a lower price than would result from a maximum with $c_2 > c_1$. Given the integer character of m_1, subsidies might be used to get the optimal p, once m_1 is set so as to get as close as possible.

A similar analysis applies to the surplus in the domestic market, but space limitations preclude analyzing it.

The general point is that subsidies as "competitive" weapons are not really of interest because they simply tip the balance in the equilibrium in a way that increases net benefits to the subsidizer. The right amount of competition will do the same thing. Rather, subsidies are a way of achieving this effect when scale economies make expanding the number of competitors costly. Moreover, this type of analysis indicates that controlling the amount of competition in the nondomestic market, especially when the country has a cost advantage in the relevant industry, is of central importance.

Finally, this argument and all the conclusions hold, independent of the structure of demand. Nothing in the preceding argument required that the elasticity of demand be independent of the price.

C. *Research and Development: Potential and Actual Spillovers*

Research and development attracts attention in multinational competitive analysis because it induces dynamic scale advantages that interact with market share. However, the distinctive feature of the *R&D* investment is that it generates information that is potentially useful to firms other than the investor. I will call this effect potential spillovers. The potential spillovers may or may not be actual: that is, they may not occur.

Spillovers dampen investment incentives. It does not, however, follow (without a considerable amount of argument) that performance is poor. The reason is that there is another effect. Spillovers increase the pace of technological progress at the industry level *for given levels of investment by firms*, because they reduce the redundancy. Put another way, an industry with high potential, but low actual, spillovers will suffer from redundancy (absent some complex interfirm transfers of technology to which I will turn later), and as a result, the dynamic technical efficiency of the industry will be impaired relative to what could be achieved. These effects, the incentive effect and the efficiency effect, work in opposite directions.

Under conditions of high potential spillovers, strictly noncooperative behavior leads to suboptimal dynamic performance, independent of the level of actual spillovers. At low actual spillovers, redundancy takes its toll, and as actual spillovers rise, incentives decline.

Market and governmental institutions adapt to the problem. The *R&D* is directly and indirectly subsidized with beneficial effects on marginal incentives. Firms, with or without the public sector as partners, jointly invest in certain parts of *R&D*. Firms also engage in voluntary transfers of technology, sometimes in the form of exchanges or cross-licensing agreements. All of the above are observable in several of the electronics industries in several countries.

Broadly speaking, the problem in high potential spillover industries is to obtain the benefits of the spillovers (i.e., make them actual) without diminishing the investment incentives.

In the multinational industry, these problems become more complex for a number of reasons. To the extent that governments are investors (directly or via subsidies) in technology in high spillover environments, the benefits will spill across boundaries unless steps are taken to interdict the transnational flows. There is, then, a second level of the free-rider problem at the country level.

Spillovers between firms in different countries may vary for legal, institutional, and policy reasons. There are obviously a number of cases here. But the main points are (*i*) that differences across countries in dealing with these externalities will lead directly to shifts in relative competitive position, (*ii*) that one-way spillovers will have the same effect, and (*iii*) that certain forms of cooperative or quasi-cooperative behavior among firms and among countries are required for dynamic efficiency.

II. Conclusions

The general point is that in certain structural contexts, policies of subsidizing and restricting access can have a significant effect on the relative competitive positions of firms in a multinational industry. These policies have been used in pursuit of competitive advantage, and they could be used strategically to prevent their use by competitors. In *R&D*, the appropriate objective is to "internalize" certain externalities by structure or policy. That objective can be pursued at the national level, but it is preferable that it be done multinationally.

The Fat-Cat Effect, The Puppy-Dog Ploy, and the Lean and Hungry Look

By DREW FUDENBERG AND JEAN TIROLE*

Let me have about me men that are fat.... *Julius Caesar*, Act 1, Sc. 2

The idea that strategic considerations may provide firms an incentive to "overinvest" in "capital" to deter the entry or expansion of rivals is by now well understood. However, in some circumstances, increased investment may be a strategic handicap, because it may reduce the incentive to respond aggressively to competitors. In such cases, firms may instead choose to maintain a "lean and hungry look," thus avoiding the "fat-cat effect." We illustrate these effects with models of investment in advertising and in *R&D*. We also provide a taxonomy of the factors which tend to favor over- and underinvestment, both to deter entry and to accommodate it. Such a classification, of course, requires a notion of what it means to overinvest; that is, we must provide a benchmark for comparison. If entry is deterred, we use a monopolist's investment as the basis for comparison. For the case of entry accommodation, we compare the incumbent's investment to that in a "precommitment" or "open-loop" equilibrium, in which the incumbent takes the entrant's actions as given and does not try to influence them through its choice of preentry investment. We flesh out the taxonomy with several additional examples.

Our advertising model was inspired by Richard Schmalensee's (1982) paper, whose results foreshadow ours. We provide an example in which an established firm will underinvest in advertising if it chooses to deter entry, because by lowering its stock of "goodwill" it establishes a credible threat to cut prices in the event of entry. Conversely, if the established firm chooses to allow entry, it will advertise heavily and become a fat cat in order to soften the entrant's pricing behavior. Thus the strategic incentives for investment depend on whether or not the incumbent chooses to deter entry. This contrasts with the previous work on strategic investment in cost-reducing machinery (Michael Spence, 1977, 1979; Avinash Dixit, 1979; our 1983a article) and in "learning by doing" (Spence, 1981; our 1983c article) in which the strategic incentives always encourage the incumbent to overaccumulate. Our *R&D* model builds on Jennifer Reinganum's (1983) observation that the "Arrow effect" (Kenneth Arrow, 1962) of an incumbent monopolist's reduced incentive to do *R&D* is robust to the threat of entry so long as the *R&D* technology is stochastic.

Our examples show that the key factors in strategic investment are whether investment makes the incumbent more or less "tough" in the post-entry game, and how the entrant reacts to tougher play by the incumbent. These two factors are the basis of our taxonomy. Jeremy Bulow et al. (1983) have independently noted the importance of the entrant's reaction. Their paper overlaps a good deal with ours.

I. Advertising and Goodwill

In our goodwill model, a customer can buy from a firm only if he is aware of its existence. To inform consumers, firms place ads in newspapers. An ad that is read informs the customer of the existence of the firm and also gives the firm's price. In the first period, only the incumbent is in the market; in the second period the entrant may

*University of California, Berkeley, CA 94707, and CERAS, Ecole Nationale des Ponts et Chaussées, 75007 Paris, France. Much of this paper is drawn from our survey (1983b). We would like to thank John Geanakoplos, Jennifer Reinganum, and Richard Schmalensee for helpful conversations. Research support from the National Science Foundation is gratefully acknowledged.

enter. The crucial assumption is that some of the customers who received an ad in the first period do not bother to read the ads in the second period, and therefore buy only from the incumbent. This captive market for the incumbent represents the incumbent's accumulation of goodwill. One could derive such captivity from a model in which rational consumers possess imperfect information about product quality, as in Schmalensee (1982), or from a model in which customers must sink firm-specific costs in learning how to consume the product.

There are two firms, an incumbent and an entrant, and a unit population of *ex ante* identical consumers. If a consumer is aware of both firms, and the incumbent charges x_1, and the entrant charges x_2, the consumer's demands for the two goods are $D^1(x_1,x_2)$ and $D^2(x_1,x_2)$, respectively. If a consumer is only aware of the incumbent (entrant), his demand is $D^1(x_1,\infty)$ and $(D^2(\infty,x_2))$. The (net of variable costs) revenue an informed consumer brings the incumbent is $R^1(x_1,x_2)$ or $R^1(x_1,\infty)$ depending on whether the consumer also knows about the entrant or not, and similarly for the entrant. We'll assume that the revenues are differentiable, quasi concave in own-prices, and they, as well as the marginal revenue, increase with the competitor's price (these are standard assumptions for price competition with differentiated goods).

To inform consumers, the firms put ads in the newspapers. An ad that is read makes the customer aware of the product and gives the price. The cost of reaching a fraction K of the population in the first period is $A(K)$, where $A(K)$ is convex for strictly positive levels of advertising, and $A(1)=\infty$.[1] There are two periods, $t=1,2$. In the first period, only the incumbent is in the market. It advertises K_1, charges the monopoly price, and makes profits $K^1\cdot R^m$. In the second period the entrant may enter.

To further simplify, we assume that all active firms will choose to cover the remaining market in the second period at cost A_2. Then assuming entry, the profits of the two firms, Π^1 and Π^2, can be written

$$(1)\quad \begin{aligned}\Pi^1 &= [-A(K_1)+K_1R^m]\\ &\quad +\delta\big[K_1R^1(x_1,\infty)\\ &\quad +(1-K_1)R^1(x_1,x_2)-A_2\big]\\ \Pi^2 &= \delta\big[(1-K_1)R^2(x_1,x_2)-A_2\big],\end{aligned}$$

where δ is the common discount factor.

In the second period, the firms simultaneously choose prices. Assuming that a Nash equilibrium for this second-stage game exists and is characterized by the first-order conditions, we have

$$(2)\quad K_1R^1_1(x^*_1,\infty) + (1-K_1)R^1_1(x^*_1,x^*_2)=0;$$

$$(3)\quad R^2_2(x^*_1,x^*_2)=0,$$

where $R^i_j\equiv\partial R^i(x_1,x_2)/\partial x_j$, and x^*_i is the equilibrium value of x_i as a function of K_1.

From equation (2), and the assumption that $R^i_{ij}>0$, we see that

$$R^1_1(x^*_1,\infty)>0>R^1_1(x^*_1,x^*_2).$$

The incumbent would like to increase its price for its captive customers, and reduce it where there is competition; but price discrimination has been assumed impossible.

Differentiating the first-order conditions, and using $R^i_{ij}>0$, we have

$$(4)\quad \begin{aligned}&\partial x^*_1/\partial K_1>0, &&\partial x^*_1/\partial x^*_2>0,\\ &\partial x^*_2/\partial K_1=0, &&\partial x^*_2/\partial x^*_1>0.\end{aligned}$$

The heart of the fat-cat effect is that $\partial x^*_1/\partial K_1>0$. As the incumbent's goodwill increases, it becomes more reluctant to match the entrant's price. The large captive market makes the incumbent a pacifistic "fat cat." This suggests that if entry is going to occur, the incumbent has an incentive to increase K_1 to "soften" the second-period equilibrium.

[1] See Gerard Butters (1977), and Gene Grossman and Carl Shapiro (1984) for examples of advertising technologies.

To formalize this intuition we first must sign the *total* derivative dx_1^*/dK_1. While one would expect increasing K_1 to increase the incumbent's equilibrium price, this is only true if firm 1's second-period reaction curve is steeper than firm 2's. This will be true if $R_{11}^1 \cdot R_{22}^2 > R_{12}^1 \cdot R_{21}^2$. If dx_1^*/dK_1 were negative the model would not exhibit the fat-cat effect.

Now we compare the incumbent's choice of K_1 in the open-loop and perfect equilibria. In the former, the incumbent takes x_2^* as given, and thus ignores the possibility of strategic investment. Setting $\partial\pi^1/\partial K_1 = 0$ in (1), we have

$$(5) \quad R^m + \delta\big(R^1(x_1^*,\infty) - R^1(x_1^*,x_2^*)\big) = A'(K_1).$$

In a perfect equilibrium, the incumbent realizes that x_2^* depends on K_1, giving first-order conditions

$$(6) \quad R^m + \delta\big(R^1(x_1^*,\infty) - R^1(x_1^*,x_2^*) + (1-K_1)R_2^1(dx_2^*/dK_1)\big) = A'(K_1).$$

As R_2^1 and dx_2^*/dK_1 are positive, for a fixed K_1 the left-hand side of (6) exceeds that of (5), so if the second-order condition corresponding to (6) is satisfied, its solution exceeds that of (5).

The fat-cat effect suggests a corollary, that the incumbent should underinvest and maintain a "lean and hungry look" to deter entry. However, while the "price effect" of increasing K_1 encourages entry, the "direct effect" of reducing the entrant's market goes the other way. To see this, note that

$$(7) \quad \Pi_K^2 = \delta\big[(1-K_1)R_1^2(dx_1^*/dK_1) - R^2\big].$$

The first term in the right-hand side of (7) is the strategic effect of K_1 on the second-period price, the second is the direct effect. One can find plausible examples of demand and advertising functions such that the indirect effect dominates. This is the case, for example, for goods which are differentiated by their location on the unit interval with linear "transportation" costs, if first-period advertising is sufficiently expensive that the incumbent's equilibrium share of the informed consumers is positive. In this case, entry deterrence requires underinvestment.

II. Technological Competition

We now develop a simple model of investment in *R&D* to illustrate the lean and hungry look, building on the work of Arrow and Reinganum. In the first period, the incumbent, firm 1, spends K_1 on capital, and then has constant average cost $\bar{c}(K_1)$. The incumbent receives the monopoly profit $V^m(\bar{c}(K_1))$ in period 1. In the second period, both the incumbent and firm 2 may do *R&D* on a new technology which allows constant average cost c. If one firm develops the innovation, it receives the monopoly value $V^m(c)$. Thus the innovation is "large" or "drastic" in Arrow's sense. If both firms develop the innovation, their profit is zero. If neither firm succeeds, then the incumbent again receives $V^m(\bar{c})$. The second-period *R&D* technology is stochastic. If firm i spends x_i on *R&D*, it obtains the new technology with probability $\mu_i(x_i)$. We assume $\mu_i'(0) = \infty$, $\mu_i' > 0$, $\mu_i'' < 0$. The total payoffs from period 2 on are

$$(9) \quad \Pi^1 = \mu_1(1-\mu_2)V^m(c) + (1-\mu_1)(1-\mu_2)V^m(\bar{c}) - x_1,$$
$$\Pi^2 = \mu_2(1-\mu_1)V^m(c) - x_2.$$

The first-order conditions for a Nash equilibrium are

$$(10) \quad \mu_1'[V^m(c) - V^m(\bar{c})](1-\mu_2) = 1,$$
$$\mu_2'V^m(c)(1-\mu_1) = 1.$$

We see that since the incumbent's gain is only the difference in the monopoly profits, it has less incentive to innovate than the entrant. This is the Arrow effect.[2] We have

[2] For large innovations, the monopoly price with the new technology is less than the average cost of the old one. Richard Gilbert and David Newbery (1982) showed

derived it here in a model with each firm's chance of succeeding independent of the other's, so that we have had to allow a nonzero probability of a tie. Reinganum's model avoids ties, because the possibilities of "success" (obtaining the patent) are not independent.

Because $\mu_i' > 0$ and $\mu_i'' < 0$, the reaction curves in (10) slope downward—the more one firm spends, the less the other wishes to. Since increasing K_1 decreases the incumbent's gain from the innovator's we expect that the strategic incentive is to reduce K_1 to play more aggressively in period 2. As in our last example, this is only true if the reaction curves are "stable," which in this case requires $\mu_1''\mu_2''(1-\mu_1)(1-\mu_2) > (\mu_1'\mu_2')^2$. This is true for example for $\mu_i(x) = \max(1, bx^{1/2})$, with b small. We conclude that to accommodate entry the incumbent has a strategic incentive to underinvest. Because K_1 has no direct effect on Π^2, we can also say that to deter entry the incumbent has an incentive to underinvest.[3]

III. Taxonomy and Conclusion

In the goodwill model the incumbent could underinvest to deter entry, while in the *R&D* model the strategic incentives always favored underinvestment. To relate these results to previous work, we next present an informal taxonomy of pre-entry strategic investment by an incumbent. In many cases, one might expect both "investment" and "production" decisions to be made post-entry. We have restricted attention to a single post-entry variable for simplicity. We should point out that this involves some loss of generality. Strategic underinvestment requires that the incumbent not be able to invest after entry, or more generally that pre- and post-entry investments are imperfectly substitutable. This was the case in both of our examples. However, if investment is in productive machinery and capital costs are linear and constant over time, then underinvestment would be ineffective, as the incumbent's post-entry investment would make up any previous restraint.

Before presenting the taxonomy, it should be acknowledged that since Schmalensee's (1983) article, several authors have independently noticed the possibility of underinvestment. J. Baldani (1983) studies the conditions leading to underinvestment in advertising. Bulow et al. present a careful treatment of two-stage games in which either production or investment takes place in the first period, with production in the second, and costs need not be separable across periods. They focus on cost minimization as the benchmark for over- and underinvestment. The starting point for the Bulow et al. paper was the observation that a firm might choose not to enter an apparently profitable market due to strategic spillovers on other product lines. This point is developed in more detail in K. Judd (1983).

Our taxonomy classifies market according to the signs of the incentives for strategic investments. Because only the incumbent has a strategic incentive, given concavity, we can unambiguously say whether the incumbent will over- or underinvest to accommodate entry (compared to the open-loop equilibrium).[4] We continue to denote the incumbent's first-period choice K_1, the post-entry decisions x_1 and x_2, and the payoffs Π^1 and Π^2. For entry deterrence there are

that for "small" innovations, because the sum of the duopoly profits is (typically) less than $\Pi^m(c)$, the incumbent loses more than the entrant gains if the entrant obtains the patent. With a deterministic *R&D* technology, the incumbent's incentive to innovate thus exceeds the entrant's, because the incumbent's current patent is certain to be superceded and thus the current profits are not "sacrificed" by the incumbent's *R&D*. Reinganum showed that with stochastic *R&D* and a small innovation, either effect can dominate. In her *R&D* model the reaction curves slope up.

[3] For small innovations the direct effect goes the other way.

[4] This does not generalize to the case in which both firms make strategic decisions. In our paper on learning by doing (1983c), we give an example in which one firm's first-period output declined in moving from the precommitment to the perfect equilibrium. The problem is that if, as expected, firm 1's output increases when it plays strategically, firm 2's strategic incentive to increase output can be outweighed by its response to firm 1's change.

TABLE 1

Slope of Reaction Curves	Investment Makes Incumbent: Tough	Soft
Upward	Case IV	Case I
	A: Puppy Dog D: Top Dog	A: Fat Cat D: Lean and Hungry
Downward	Case III	Case II
	A: Top Dog	A: Lean and Hungry
	A: Top Dog	A: Lean and Hungry

Note: A = Accommodate entry; D = Deter entry.

two effects, as we noted before: the "direct effect" $\partial\Pi^2/\partial K_1$, and the "strategic effect" $\partial\Pi^2/\partial x_1^* \cdot \partial x_1^*/\partial K_1$. We saw in the goodwill case that these two effects had opposite signs, and so the overall incentives were ambiguous. In all the rest of our examples, these two effects have the same sign.

In Table 1, first the entry-accommodating strategy and then the entry-deterring one is given. The fat-cat strategy is overinvestment that accommodates entry by committing the incumbent to play less aggressively post-entry. The lean and hungry strategy is under-investment to be tougher. The top dog strategy is overinvestment to be tough; this is the familiar result of Spence and Dixit.

Last, the puppy-dog strategy is *underinvestment* that accommodates entry by turning the incumbent into a small, friendly, nonaggressive puppy dog. This strategy is desirable if investment makes the incumbent tougher, and the second-period reaction curves slope up.

One final caveat: the classification in Table 1 depends as previously on the second-period Nash equilibria being "stable," so that changing K_1 has the intuitive effect on x_2^*.

Our goodwill model is an example of Case I: goodwill makes the incumbent soft, and the second-period reaction curves slope up. The *R&D* model illustrates Case II. Case III is the "classic" case for investing in productive machinery and "learning by doing" (Spence, 1981; our paper, 1983c) with quantity competition. Case IV results from either of these models with price competition (Bulow et al.; our paper, 1983b; Judith Gelman and Steven Salop, 1983). A more novel example of the puppy-dog ploy arises in the P. Milgrom and J. Roberts (1982) model of limit pricing under incomplete information, if we remove their assumption that the established firm's cost is revealed once the entrant decides to enter, and replace quantity with price as the strategic variable. To accommodate entry, the incumbent then prefers the entrant to believe that the incumbent's costs are relatively high.

We conclude with two warnings. First, one key ingredient of our taxonomy is the slope of the second-period reaction curves. In many of our examples, downward slopes correspond to quantity competition and upward slopes to competition in prices.[5] These examples are potentially misleading. We do not intend to revive the Cournot vs. Bertrand argument. As David Kreps and José Scheinkman (1983) have shown, "Quantity Precommitment and Bertrand Competition Yield Cournot Outcomes." Thus, "price competition" and "quantity competition" should not be interpreted as referring to the variable chosen by firms in the second stage, but rather as two different reduced forms for the determination of both prices and outputs. Second, our restriction to a single post-entry stage eliminates many important strategic interactions. As our 1983a paper shows, such interactions may reverse the over- or under-investment results of two-stage models.

[5] Bulow et al. point out that while these are the "normal" cases, it is possible, for example, for reaction curves to slope up in quantity competition.

REFERENCES

Arrow, Kenneth, "Economic Welfare and the Allocation of Resources to Innovation," in R. Nelson, ed., *The Rate and Direction of Economic Activity*, New York: National Bureau of Economic Research, 1962.

Baldani, J., "Strategic Advertising and Credible Entry Deterrence Policies," mimeo., Colgate University, 1983.

Bulow, J., Geanakoplos, J. and Klemperer, P., "Multimarket Oligopoly," Stanford Business School R. P. 696, 1983.

Butters, Gerard, "Equilibrium Distributions of Sales and Advertising Prices," *Review of Economic Studies*, October 1977, *44*, 465–96.

Dixit, A., "A Model of Duopoly Suggesting a Theory of Entry Barriers," *Bell Journal of Economics*, Spring 1979, *10*, 20–32.

Fudenberg, D. and Tirole, J., (1983a) "Capital as a Commitment: Strategic Investment to Deter Mobility," *Journal of Economic Theory*, December 1983, *31*, 227–50.

______ **and** ______, (1983b) "Dynamic Models of Oligopoly," IMSSS T. R. 428, Stanford University, 1983.

______ **and** ______, (1983c) "Learning by Doing and Market Performance," *Bell Journal of Economics*, Autumn 1983, *14*, 522–30.

Gelman, J. and Salop, S., "Judo Economics," mimeo., George Washington University, 1982.

Gilbert, R. and Newbery, D., "Preemptive Patenting and the Persistence of Monopoly," *American Economic Review*, June 1982, *72*, 514–26.

Grossman, G. and Shapiro, C., "Informative Advertising with Differentiated Goods," *Review of Economic Studies*, January 1984, *51*, 63–82.

Judd, K., "Credible Spatial Preemption," MEDS D. P. 577, Northwestern University, 1983.

Kreps, D. and Scheinkman, J., "Quantity Precommitment and Bertrand Competition Yield Cournot Outcomes," mimeo., University of Chicago, 1983.

Milgrom, P. and Roberts, J., "Limit Pricing and Entry under Incomplete Information," *Econometrica*, 1982, *50*, 443–60.

Reinganum, Jennifer, "Uncertain Innovation and the Persistence of Monopoly," *American Economic Review*, September 1983, *73*, 741–48.

Schmalensee, Richard, "Product Differentiation Advantages of Pioneering Brands," *American Economic Review*, June 1982, *72*, 349–65.

______, "Advertising and Entry Deterrence: An Exploratory Model," *Journal of Political Economy*, August 1983, *90*, 636–53.

Spence, A. Michael, "Entry, Capacity, Investment, and Oligopolistic Pricing," *Bell Journal of Economics*, Autumn 1977, *8*, 534–44.

______, "Investment Strategy and Growth in a New Market," *Bell Journal of Economics*, Spring 1979, *10*, 1–19.

______, "The Learning Curve and Competition," *Bell Journal of Economics*, Spring 1981, *12*, 49–70.

ANTITRUST POLICY AND STRATEGIC BEHAVIOR

Noisy Advertising and the Predation Rule in Antitrust Analysis

By JOHN C. HILKE AND PHILIP B. NELSON*

Can advertising by dominant firms pose a predatory threat? Theoretically, the answer clearly is "yes." However, empirical evidence indicating the importance of this theoretical finding is largely missing. Economists' empirical studies of advertising, which typically focus on the question of whether high levels of advertising are indicative of the presence of a barrier to entry or signal that advertising information can substitute for consumption experience, usually employ average advertising levels that hide strategically focused price cuts or changes in short-term advertising rates.

I. Predatory Advertising: Theory and Policy Implications

Here we review two theories that describe how dominant firms can employ advertising to predate[1] against an entrant and relate these theories to the advertising strategy used by a dominant consumer goods firm that responded to the geographic expansion of a large rival.

THEORY 1: *The basic premise of recent work on raising rivals' costs*[2] *is that even if firms initially are equally efficient producers (although they have different cost structures), additional expenditures on some items by one firm can cause more than proportionate increases in the costs of its competitors.*

Applying this model to advertising, strategic advertising behavior by a dominant firm that is responding to entry *could* be of concern if costs per advertisement are higher for a small seller trying to expand, and if advertising levels of the dominant firm have to be matched in some fashion by the smaller seller. Alternatively, it might be that advertising is normally a cost effective way of seeking new customers, but increased advertising by the dominant firm forces the entrant to shift to less-cost-effective techniques. Specifically, if consumers use information both from advertising and from experience, actions by the dominant firm which disrupt the flow of information to consumers from advertising can give the dominant firm an advantage, since consumers will have more experience with the dominant firm's products than with the entrant's products.

Psychology research on information overload is consistent with the view of Jacob Jacoby that additional advertising may disrupt all firms' ability to communicate through advertising: "there are finite limits to the ability of human beings to assimilate and process information during any given unit of time. Once these limits are surpassed, the system is said to be 'overloaded' and human performance (including decision-making) becomes confused, less accurate, and less effective" (1977, p. 569).

In addition, memory research, which has focused both on how much information is registered and on how much of the registered information can be retrieved, has found that the presentation of similar information reduces memory for the original items of information. Alan Baddeley summarizes this research, noting "There is a good deal of

*Staff Economist and Assistant Director for Competition Analysis, respectively, Federal Trade Commission, Bureau of Economics, 6th and Pennsylvania Ave., NW, Washington, D.C. 20580.

[1] By "predate," we mean a response to a rival that sacrifices part of the profit that could be earned under competitive conditions in order to gain exit or other long term advantage over the rival and consequently to gain additional monopoly profit. This is a paraphrase of Janusz Ordover and Robert Willig (1981).

[2] See Steven Salop and David Scheffman (1983) and Judith Gelman and Salop (1982).

evidence for such a position [interference influences forgetting but is not necessarily the only factor], particularly from a vast range of studies which have shown that the amount retained is reduced if the learning of other material has occurred during the retention interval. Forgetting is particularly marked if the interfering material is similar to that originally learned" (1976, p. 62). Such interference can be either retroactive, learning a second set of information reduces recall of the first set of information, or proactive, learning an earlier set of information reduces recall of a second set of information.[3] These effects are likely to be more pronounced in incidental memory situations.[4]

THEORY 2: *An entrant may be especially vulnerable to its competitor's activities during its initial period of operation.*

The time dependency of the vulnerability of the entrant distinguishes this theory from the previous "cost-raising" theory. Furthermore, the asymmetry that supports predation does not require the firms to have different costs as does Theory 1: the firms may have identical costs at similar points in their life cycles, *ceteris paribus*. However, according to the "infant firm" theory, actions by the dominant firm during a competitor's entry may shift the market environment so that the entrant's cost trajectory is changed. Specifically, the dominant firm's expenditure of a dollar more on advertising during the entry period may force the entrant to spend $5 in later recuperative advertising that wouldn't have been necessary absent the action by the dominant firm.

Why should an entrant be particularly vulnerable? While there probably are many reasons, two examples should suffice. First, retailer restocking decisions are often based on comparing actual initial sales to expected sales of the new product, since there is no long purchase history to depend upon. As a result, an entrant's distribution network may be more vulnerable to the promotional campaigns of the dominant firm during entry than it will be later. Similarly, if consumer openness to experimenting with a new product is greatest when it is first introduced, a dominant firm's promotional campaign during a rival's entry may be particularly detrimental to the entrant's long-run costs of providing information to consumers.

Again note that, unlike the cost-raising model which can involve stops and starts in the small firm's or entrant's efforts, the infant firm theory emphasizes the need for continuity in an introductory campaign.

With only a few exceptions (see Hilke, 1980), discussions of appropriate policies for detecting predation have focused on pricing behavior, not on alternative tactics such as increased advertising. At least theoretically, this appears to be a mistake. There is no reason to believe that a rule designed to identify predatory pricing will be appropriate for detecting predatory advertising. As a result, laws designed to deter predatory pricing may simply lead predators to employ predatory advertising instead. And, because predatory advertising consumes resources, it may increase the social losses from predation.

II. Observations of Advertising Strategies in Action

The theoretical possibility of predatory advertising is not sufficient to warrant policy concern if firms do not or would not engage in activities that might have the predicted consequences. Here, as a first step, we report findings from our investigation of the advertising strategies used by a leading firm in a consumer goods industry when responding to the geographic expansion of a rival. The case is FTC Case D-9085, In The Matter Of General Foods Corporation. General Foods' expanding rival was Folger Coffee Co., a division of Proctor and Gamble.

Since some short-run profit-maximizing considerations, as well as predation, might motivate a leading firm to increase advertising, the case study should be viewed as an effort to get a glimpse at strategic advertising

[3] See Baddeley (chs. 1–5) and Donald Norman (1976, chs. 5–6) for comprehensive reviews of the memory literature. See also James Bettman (1979, especially pp. 37–39).

[4] See Baddeley, pp. 46–49, and Bettman, pp. 45–46.

behavior in order to test the relevance of the earlier theories. Our hope is that results that are at least consistent with the two theories will lead others to examine similar situations and enlarge our understanding of entry and advertising activities. A primary objective for these inquiries should be to develop means of effectively identifying classes of strategic behavior that should to be deemed "predatory."

During the 1970's Folger (brand F) started marketing in areas of the East where it had not previously operated and where Maxwell House (brand M) had approximately 4 to 1 share positions relative to the second largest brands; 44 vs. 11 percent on average. Strategy and marketing documents of M and F, as well as testimony of market participants, indicate that there were some variations in M's (and F's) tactics during the incremental expansion by F. However, there were major consistent elements in M's advertising strategy:[5]

Brand M's documents indicate that it substantially increased its advertising just prior to the entry dates in order to 1) encourage consumers to stock up on M and make them less likely to pay attention to F's ads and to reduce the likelihood that they would try the new brand; 2) absorb retailers' warehouse space that might otherwise be available for stocking the new brand; and 3) coordinate with pricing and other promotional vehicles to minimize retailer and consumer interest in the entrant.

Brand M abandoned traditional levels of advertising activity and determined the amount of its advertising on the basis of the amount of advertising done by F. Its target was to out-advertise the entrant by 50 percent in terms of advertising exposures. Dominance of the advertising media was expected to discourage initial trial of F and to disrupt the emergence of repeat purchase behavior. The belief was frequently expressed that F's best opportunity to establish itself with consumers and retailers would occur while it was regarded as new. Since this period of novelty would be finite, M could contemplate advertising and other activities that it could not sustain in the long run, but which would reduce the initial market penetration of the new entrant to the entrant's long-run disadvantage.

Brand M developed advertising similar to the advertising used by the entrant to reduce the apparent novelty of the entrant's ads. M expected that by reducing this novelty, consumer reactions to the entrant's ads would be reduced. To the extent that consumers were confused by this similarity it could redound to the benefit of M because of greater consumer familiarity with the leading brand.

When one compares the earlier discussions of predatory advertising theories to M's advertising strategies, several connections appear. M's documents mention using higher levels of advertising to increase F's costs. There is also some indication that advertising can be substituted for other forms of competition, if these avenues were for some reason constrained. But the preeminent assumption and basis for the actual strategies had to do with timing and the infant enterprise theory.

In addition to strategy considerations, which to some extent reveal the participants' views of market characteristics, the case documents provide data that allow some, albeit limited, direct empirical testing of the impact of M's strategies. Table 1 reports advertising and brand awareness data for five sequential entry areas by month or quarter.

The question we would like to answer with this data is whether or not M's advertising strategies proved to be a cost effective way to stunt F's growth in the East. Unfortunately, too many things that reasonably might enter into determining the levels of recall and brand awareness were not recorded in the available data to allow conclusive testing. But enough data are available to see if the results of M's increased advertising are consistent with its own strategic thinking.

In market *A*, brand M increased advertising at the time of entry and built to higher levels for several months thereafter before dropping back. In the other TV markets, M undertook the heaviest levels of advertising just before and during F's initial advertising.

[5] See the public version of *Complaint Counsel's Proposed Findings of Fact*, Vol. I, Section VII. Share figures are shown on pp. 200–03.

TABLE 1—AVERAGE WEEKLY ADVERTISING EXPOSURES AND PERCENT OF CONSUMERS WITH UNAIDED RECALL OF A BRAND'S ADVERTISEMENTS[a]

Area–Month	M	F
A–before	260(NA)	NA
A–1	240(29)	155(47)
A–2	270(40)	160(47)
A–3	330(35)	165(51)
A–4	295(47)	140(57)
A–5	295(39)	145(57)
A–6	160(49)	125(50)
A–9	70(47)	125(48)
A–12	155(51)	180(44)
B–before	250(NA)	88(NA)
B–1	170(68)	130(44)
B–2	180(67)	175(52)
B–3	155(68)	190(58)
C–before	160(NA)	NA
C–1	260(51)	175(24)
C–2	210(41)	175(45)
C–3	190(43)	170(42)
C–6	225(51)	145(29)
D–before	230(NA)	NA
D–1	220(69)	180(46)
D–2	190(67)	180(62)
D–3	180(68)	180(59)
E–before	NA(62)	NA(4)
E–3	240(64)	155(58)
E–6	170(66)	145(60)

Sources: Brand M's documents, Commission Exhibits CX449, CX450, CX462, CX770, and CX773.

[a] The first number in the columns is the measure of advertising intensity. It is the average weekly exposures to advertisements of a given brand for female TV viewers as reported to M. The number shown in parentheses is the percent of regular coffee drinkers who recalled hearing advertisements for a brand without prompting from the interviewer. NA = Not Available.

Brand M then reduced advertising briefly, before returning to a high level. This difference in tactics allows one to test whether M may have gained from responding with a wave pattern of advertising. If the wave pattern is more effective, we should observe a higher relative advertising recall for F in area *A* than in the subsequent entries. In fact this is what is found. Despite having higher ratios of F's advertisements to M's advertisements in the later markets (*B* to *E*), the average ratio of F's recall to M's recall is lower in the later markets: 1.24 vs. .81.

Brand M also changed its advertising copy to align much more closely with F's campaign during the period, so it would be useful to test whether this had an effect. Brand M's new campaign appeared between observations 3 and 4 in TV market *A*. If the new campaign produced the reduction in the perceived novelty of F's ads that M anticipated, then F's relative awareness levels in subsequent markets should be lower. Unfortunately, the decisions to change copy and to adopt the wave pattern advertising were part of the same bundle of strategy decisions. Thus, while the results are consistent with the advertising copy theory, they are confounded. However, we might also expect to find an effect within the market *A* data before and after the switch in advertising tactics. The result of comparing observations *A*–1 through *A*–3 with *A*–4 through *A*–8 is consistent with the advertising copy hypothesis: the average ratio of F's recall to M's recall is 1.43 up to observation *A*–3 and 1.12 after that. However, this result is also consistent with the expectation that there would be strong initial interest in M's program and that the novelty of F's program would have begun to fade by then.[6]

III. Conclusions

We derive two main conclusions from our review of concerns about strategic advertising activity. First, an important step toward substantiating a concern about strategic advertising activity is increasing our understanding of the time sequence of consumer and retailer information acquisition about new products. If there is a critical entry

[6] Other relationships in the data are also consistent with an interference or noise view of the dominant firm's advertising. There are no instances in which recall of the entrant's (F's) ads increased while the entrant's ads decreased and M's ads increased. A model in which changes in recall of the entrant's ads are predicted simply by changes in the entrant's advertising program does not work well. Such a model predicts correctly 7 times out of 15. A model predicting changes in recall of F's ads on the basis of changes in the relative advertising levels of M and F produces 10 successes out of 15 trials (significant at .1509 using the binomial distribution). A model predicting changes in relative recall on the basis of relative changes in advertising levels produces 12 successes (significant at .0176 using the binomial distribution).

period, then consideration of advertising and indeed other strategies during that period deserve attention. Second, clearer findings would be obtainable if more variables could be controlled and monitored. However, in real marketing situations, actors have strong incentives to use a whole portfolio of tactics simultaneously. Separating the effects is extremely difficult. Thus, the problems of separating effects and then distinguishing between objectionable and nonobjectionable advertising responses to entry remain major obstacles to empirical understanding of advertising dynamics in entry and to more refined policy consideration of this phenomenon.

REFERENCES

Baddeley, Alan, *The Psychology of Memory,* New York: Basic Books, 1976.

Bettman, James, "Memory Factors In Consumer Choice: A Review," *Journal of Marketing*, Spring 1979, *43*, 37–53.

Gelman, Judith R. and Salop, Steven C., "Judo Economics, Entrant Advantages, and the Great Airline Coupon Wars," FTC Bureau of Economics Working Paper No. 58, June 1982.

Hilke, John C., "Advertising Predation and the Areeda-Turner and Williamson Rules," *Journal of Reprints for Antitrust Law and Economics*, No. 1, 1980, *10*, 367–98.

Jacoby, Jacob, "Information Load and Decision Quality: Some Contested Issues," *Journal of Marketing Research*, November 1977, *14*, 569–73.

Norman, Donald, *Memory and Attention*, New York: Wiley & Sons, 1976.

Ordover, Janusz A. and Willig, Robert D., "An Economic Definition of Predatory Product Innovation," in Steven C. Salop, ed., *Strategy, Predation and Antitrust Analysis*, Washington: Federal Trade Commission, 1981.

Salop, Steven C. and Scheffman, David T., "Raising Rivals' Cost," *American Economic Review Proceedings*, May 1983, *73*, 267–71.

Strategic Behavior and Antitrust Analysis

By William S. Comanor and H. E. Frech III*

The behavior of firms in oligopolistic markets has long posed a major conundrum to economists. That firms respond to the presence of mutual interdependence is undeniable, which indicates only that they behave strategically. In this paper, we examine some forms of strategic behavior which have received recent attention. In addition, we consider the implications for antitrust analysis of this conduct.

I. Strategic Behavior, Entry Deterrence, and Predatory Conduct

Firm behavior in oligopolistic markets is typically strategic. Indeed, what one generally expects from the recognition of mutual interdependence in oligopoly is that the expected reactions of rivals is accounted for in all market decisions. Only if one posits Nash behavior generally is this interdependence ignored and firms behave nonstrategically.

That strategic considerations are typically unavoidable is seen from the general first-order conditions for profit maximization for a duopolist:

$$\frac{d\pi_1}{dX_1} = \frac{\partial \pi_1}{\partial X_1} + \frac{\partial \pi_1}{\partial X_2}\frac{\partial X_2}{\partial X_1} = 0$$

where X_1 and X_2 are any action taken by the original firm and rival firm, respectively. The first term indicates the direct effect of the action and the second the indirect effect, which is the source of strategic decisions. Only when the second term is ignored are strategic considerations avoided.

The introduction of strategic considerations into firm decisions is thereby not caused by abandoning the assumption of profit-maximizing behavior. Where indirect effects are substantial, profit maximization requires that they be taken into account. But strategic behavior has another facet as well. Not only does it suggest that the firm accounts for the reactions of its rivals, but also it encompasses conduct specifically designed to influence a rival. These statements, however, differ only in terms of the primary purpose or intent of the action taken.

Actions which represent either predatory conduct or entry deterrence are strategic in nature, although not all strategic behavior takes these forms. The former is designed to lead existing rivals to certain choices, while the latter emphasizes the impact on prospective entrants.

Another distinction hinges on the investment nature of predatory conduct, or the pattern of effects over time, that is not necessarily present in entry deterrence. More specifically, predatory conduct are actions taken by a firm at some current cost to itself which are designed specifically to lower a rival's profits and induce him either to exit or compete less vigorously.

In this context, F. M. Scherer explicitly recognizes the "short-run profit sacrifices associated with predatory pricing" (1980, p. 339). More generally, including nonprice actions, Janusz Ordover and Robert Willig write that "predation should be defined as response to a rival that sacrifices part of the profit that could be earned under competitive circumstances, were the rival to remain viable, in order to induce exit and gain subsequent monopoly profit" (1981, p. 302).

As currently used, the concept of predation requires two distinct elements. First, predatory conduct is designed specifically to affect a rival by influencing the parameters on which his optimal decisions must rest. There must therefore be a prey. In the extreme, these parameters are affected in such a manner that his best action is to go out of business. What is essential is that the major gains from such conduct result from the indirect effects, or the second term of the expression above, although there may be direct effects as well.

*University of California, Santa Barbara, CA 93106.

The second part of the story is that there must be some element of intertemporal sacrifice. What is required here is that the predator must sacrifice at least a portion of his profits in the short run to impose harm on his prey. The objective is to convince the rival to change its behavior.

The act of predation is designed specifically to insure that the rival will believe the firm's commitment to its policy. In this manner, all predation is inherently related to the making of threats and promises. The actual carrying out of any predatory action is designed solely to make the underlying threat believed. Indeed, the optimal predatory act is a threat or commitment that is believed but never carried out.

Since the primary function of predation is to communicate to and thereby influence a rival's behavior, it may not be rational in the short run. Actions are taken which reduce the firm's short-run profits in the hope that profits will be much higher once the rival's behavior is changed, perhaps by leaving the market. While this conduct may be profitable from a long-run viewpoint, it may include a commitment to narrowly irrational acts for certain periods of time, in that profits are lower than they could be. It is this association of predation with non-profit-maximizing conduct in the short run which has lead to the view that predatory conduct will not occur.[1] One cannot explain this behavior by imposing strict standards of profit maximization at each point in time.

While most discussions of predatory conduct have emphasized the effect of a firm's actions on the effective price which rivals can charge for their products, and indeed undercutting a rival's price is the classic story of predation, Steven Salop and David Scheffman in a recent important paper (1983), call attention to behavior designed specifically to raise rivals' costs. Some examples are mentioned which have the common theme that by taking certain actions which involve increased costs, a rival's costs are increased even more.

Such actions are necessarily strategic. Higher profits result directly from the higher costs imposed on rivals. But it is not predatory under the criteria suggested above. The critical element of a short-run sacrifice of profits for higher earnings later on is missing. Higher profits are immediately achieved. Salop and Scheffman write:

> Raising rivals' costs has obvious advantages over predatory pricing.... [It] can be profitable even if the rival does not exit from the market. Nor is it necessary to sacrifice profits in the short run for "speculative and indeterminate" profits in the long run... Because these strategies do not require a sacrifice of profits in the short run, but allow profits to be increased immediately, the would-be predator has every incentive to carry out its threats. [p. 267]

Since the proposed actions are narrowly rational, they are directly carried out, and there is no element of a threat conveyed. At the same time, such actions may make other forms of predatory conduct less costly to carry out.

Salop and Scheffman suggest that this strategy can be directed against prospective entrants as well as existing firms. So long as actions are taken which increase costs solely for the relative disadvantages under which new entrants are placed, they represent strategic entry deterrence.

While entry deterrent policies are inherently strategic, they may or may not be predatory. A favorable government product standard may be an example of nonpredatory deterrence. Both established firms and new entrants believe this standard will remain indefinitely, and therefore any relative cost effects will also remain. There is no issue of communicating anything to a rival, nor will policies be changed if the rival alters his actions.

On the other hand, the threat to increase output in the face of entry represents a predatory form of entry deterrence. Here, the established firm makes this message known to prospective rivals in the hope that it will never have to carry through with this threat.

[1] For example, see John McGee's (1958) path-breaking attack on the then-traditional view that the early Standard Oil Company used predatory pricing to achieve a monopoly position. While the empirical evidence may be correct, the theoretical argument that predation does not pay rests on the inappropriate assumption of narrow rationality.

While not rational in the short run, since it will sacrifice profits by doing so, carrying out the threat may lead to higher profits in the long run.

The classic problem with predatory entry deterrence, as with other forms of predatory conduct, is that of being believed by the prey. It is inherently difficult for the original firm to bind his own later actions with the threat. Once entry occurs, what is to keep the firm from reverting to narrowly rational decisions? Nonpredatory forms of entry deterrence avoid this problem.

The best known example of a predatory entry deterrent policy is the limit-entry price behavior proposed originally by Joe Bain (1956, p. 97). Here the established firm sets price and output such that entry is not profitable if this output level is maintained. And the firm's commitment to maintain quantity in the face of entry is the essential part of this behavior. This policy is predatory for it represents a departure from short-run profit-maximizing behavior. The latter requires that existing firms accommodate entry to some extent; and reduce output once entry has taken place. Only the long-run gains from maintaining the firm's existing market position make it advantageous to bear the short-run costs in terms of profits foregone.

In regard to limit-entry pricing as well as other forms of entry deterrence, it is the willingness to bear costs rather than their actual presence which is critical to the potential entrants' decisions. A more effective deterrent than even the pledge to hold output constant in the face of entry would be a commitment to expand output, which would drive the market price below the entrant's variable costs and ensure losses. But if this type of commitment can be made, why should the firm originally forego maximum profits by setting price below the level at which marginal revenues equal marginal costs?

Consider an established firm that sets a high price originally, but then reduces it to the limit-entry value once entry occurs. Whether or not this conduct represents a sacrifice in short-run profits depends on the nature of strategic interactions among existing rivals. If the optimal short-run price following entry exceeds the limit entry price, then the price cut to exclude entrants embodies a decline in immediate profits and is, therefore, predatory. If, however, the firm's optimal price following entry just equals the limit-entry value, then there is no decline in profits and no element of predation.

An interesting distinction between predatory entry deterrence and other forms of predatory conduct is that entry may be deterred merely by the commitment to take certain actions and bear certain costs, while forcing the exit of existing firms more often requires the actual adoption of the required policies. Mere threats may not lead an existing firm to depart because of his investment in industry-specific capital.

An important conclusion from this discussion is that both strategic entry deterrence and predatory conduct require an underlying asymmetry between the original firm and the rival for its success. This asymmetry, however, may simply be that one firm is in a position to make a commitment before its rival can do so. After that, the commitment is a datum, and the rival's actions are constrained by it.[2] As a practical matter, this may mean that the original firm in the market, or the largest, may have an important advantage.

II. Implications for Antitrust

A fundamental conclusion from this discussion is that the modern emphasis on strategic behavior represents a rejection of the easy application of conventional price theory which rests on narrowly rational, short-run objectives. Indeed, as John McGee's early paper pointed out, predatory conduct is an unlikely occurrence in such instances. Its presence can be explained only when strategic concerns are emphasized.

For this reason, the development of suitable antitrust rules to deal with such problems can also not rest on the implications of standard microeconomic analysis. Yet this was precisely the attempt made in the pathbreaking paper by Phillip Areeda and Donald Turner (1975). Their object was to provide

[2] See Earl Thompson and Roger Faith (1981), especially p. 368.

an antitrust diagnosis of the problem of predatory behavior. While that paper served as the major stimulus for further work in this area, it stumbled over its attempt to deal with this problem in a nonstrategic context. Indeed, the major critiques of that work by Scherer (1976) and Oliver Williamson (1977) emphasized the importance of strategic considerations. In our judgment, no suitable policy direction can be designed without giving full recognition to the strategic elements which necessarily underlie this behavior.

Unfortunately, when allegedly predatory acts as well as other types of strategic behavior are placed in this broader context, it becomes particularly difficult to distinguish them from procompetitive conduct. A firm may merely be attempting to stimulate the demand for its products, either through setting low prices or some other means, so that the actions by themselves do not provide a distinguishing mark. What becomes important in any appraisal of this conduct, is both the purpose for which the actions are taken, or the intent of the particular decision makers, as well as the effect of the actions on both established and potential rivals. As a result, simple rules may not help to determine the presence of predation. A detailed investigation of the purpose and effects of specific acts under the Rule of Reason may therefore be necessary.

This type of antitrust recommendation, made earlier by Scherer (1976, p. 890), rests on a strategic approach to the economics of market behavior. We note its similarity to an older, more traditional form of antitrust analysis that rests more strongly on legal than economic analysis. Lawrence Sullivan, for example, argues that we should look to purpose or intent in any examination of the antitrust consequences of particular events. While emphasizing that lawyers, judges and juries deal competently with such concerns, he writes that: "purpose may be the last factor about which an economist would ask when analyzing market conduct" (1977, p. 195). While this statement may be correct when economic analysis is limited to standard price theory, it is hardly so when the broader concerns of strategic behavior are taken into account. Purpose and intent become important elements in determining the competitive consequences of various strategic actions. A major implication of the recent emphasis on strategic behavior may be a new acceptance of the role of these considerations in antitrust analysis.

In our 1983 paper, we examine the competitive implications of exclusive dealing arrangements. In some market situations, we find that the adoption of these arrangements may lead to limit entry prices which exceed competitive levels. While these actions may have anticompetitive effects, they cannot be regarded as predatory from the criteria suggested earlier. Indeed, they are more in the tradition of the Salop-Scheffman model which relates to firm efforts designed specifically to raise rivals' costs.

Note that so long as no element of predation is alleged, there is less of a role for an examination of intent. All that is relevant is the impact of a system of exclusive dealing arrangements. The issue of purpose or intent becomes critical only when the alleged anticompetitive conduct directly concerns how firms communicate threats and promises to their rivals and with the impact of this communication.

In the earlier discussion, we note that predatory conduct may be marked as narrowly irrational. But an antitrust prohibition of all such conduct would seem too broad. Even ignoring difficult problems of detection and enforcement, we should ask whether this type of prohibition serves the economic objectives of consumer welfare. Note that such a prohibition would encompass limit entry pricing strategies, since these are effectively predatory, and yet there is some doubt as to whether any alternate pricing regime would lie more in the consumer's interests.

A tempting approach to this conundrum is to stress the importance of monopoly rents. Since predation is designed to garner these rents, a prohibition of all predatory conduct would tend to lower them. If this is our objective, then the appropriate conclusion might well be to limit predatory conduct to the greatest extent possible.

Unfortunately, that conclusion rests on false foundations. Actions which minimize the present value of monopoly rents do not necessarily minimize the welfare loss to consumers. While many factors lead to a difference between monopoly rents and welfare losses, a critical one is that monopoly rents are increased or limited by actions taken at the margin while welfare gains or losses are essentially intramarginal. Another factor is the possibility that predation may limit the costs incurred to achieve particular market positions so that welfare losses are again not represented by the monopoly rents achieved. For both reasons, there is no necessary correspondence between monopoly rents and the relevant normative criteria.

Our growing understanding of the strategic conduct of firms implies a recognition of the enormous diversity of business behavior. Not only may simple rules not be useful, but also there may be advantages from returning to the traditional legal concepts of purpose or intent. Particularly where predatory conduct is alleged, such concerns may be essential in making appropriate antitrust judgments. At the same time, much strategic behavior may be nonpredatory in character, where questions of purpose or intent are less important. The new economic studies of strategic behavior place these judicial concerns on a firmer intellectual footing. As a result, the new antitrust conclusions that follow may not be new at all.

REFERENCES

Areeda, Phillip and Turner, Donald F., "Predatory Pricing and Related Practices Under Section II of the Sherman Act," *Harvard Law Review*, February 1975, *88*, 697–733.

Bain, Joe, *Barriers to New Competition*, Cambridge: Harvard University Press, 1956.

Comanor, W. S. and Frech III, H. E., "The Competitive Effects of Vertical Agreements," Working Paper No. 223, Department of Economics, University of California-Santa Barbara, June 1983.

McGee, John S., "Predatory Price Cutting: The Standard Oil (N.J.) Case," *Journal of Law and Economics*, October 1958, *1*, 137–169.

Ordover, Janusz and Willig, Robert D., "An Economic Definition of Predatory Product Innovation," in Steven C. Salop, ed., *Strategy, Predation, and Antitrust Analysis*, Washington: Federal Trade Commission, September 1981.

Salop, Steven C. and Scheffman, David P., "Raising Rival's Costs," *American Economic Review Proceedings*, May 1983, *73*, 267–71.

Scherer, F. M., *Industrial Market Structure and Economic Performance*, 2d ed., Chicago: Rand McNally, 1980.

______, "Predatory Pricing and the Sherman Act: A Comment," *Harvard Law Review*, March 1976, *89*, 869–90.

Sullivan, Lawrence A., *Handbook of the Law of Antitrust*, St. Paul: West Publishing, 1977.

Thompson, Earl A. and Faith, Roger I., "A Pure Theory of Strategic Behavior and Social Institutions," *American Economic Review*, June 1981, *71*, 366–80.

Williamson, Oliver E., "Predatory Pricing: A Strategic and Welfare Analysis," *Yale Law Journal*, December 1977, *87*, 284–340.

Commodity Bundling

By Robert E. Dansby and Cecilia Conrad*

The courts and antitrust enforcement agencies recognize that some types of commodity bundling activities serve the public interest, while others inflict substantial harm. The Clayton Act prescribes limits on the use of bundling for anticompetitive purposes. However, the case law which has developed from its application gives a sometimes conflicting set of criteria for judging permissible tie-in arrangements. The difficulties can be seen, for example, in the discussion of patent tie-ins in David Lewis (1982). What criteria should we use in setting the boundaries of lawful bundling? Will prohibitions against selected bundling activities be effective? Or will effective enforcement require the use of injunctive pricing restraints? These questions motivate the research begun in this note.

The traditional commodity bundling literature provides the foundation for our analysis. George Stigler (1968) gives insights regarding the use of commodity bundling as an anticompetitive mechanism. William Adams and Janet Yellen (1976) analyze profitability of commodity bundling by a two-product monopolist. Richard Schmalensee (1982) examines the optimal bundling strategies of a firm with a single product monopoly. Thomas Palfrey (1983) shows that the social harm caused by certain auction bundling strategies diminishes as the number of consumers, who may differ in their probable valuation of alternative bundles, increases. In this literature, a bundle is valued at the sum of the value of the bundle's component parts. This note introduces diversity of consumer bundle preferences as one element of a more general commodity bundling theory. This diversity can give firms strong incentives to bundle, even if the firm has no monopoly power. Thus, our analysis gives additional reasons, beyond those discussed in Richard Craswell (1982), that bundling may benefit both buyers and sellers. Our model also suggests criteria for characterizing bundling activities that, for these reasons, should be legal, see Michael Ross (1982).

I. Demands for Commodity Bundles and Components

This section discusses the commodity options available to consumers and the aggregate demands that result from their self-selective decisions. Assume that two commodities X and Y, are sold at prices P_x and P_y, respectively. A bundle B, consisting of one unit of X and of Y, may also be offered at price P_b. Each consumer's utility function is assumed to have the following properties: 1) the marginal utility of a second unit of either X or Y is zero; 2) the utility of consuming X alone is $x = U(X=1,0)$, while the utility of consuming Y alone is $y = U(0, Y=1)$. Each consumer's utility, b, from consumption of a bundle B is a function of the utility of consuming X *and* Y separately, that is $b \equiv b(x, y)$, which is continuous and differentiable. The utility structure standardly used in the bundling literature is $b(x, y) = x + y$. We permit b to be either orthogonally subadditive, that is, $b(x, y) < x + y$, or orthogonally superadditive, that is, $b(x, y) > x + y$. The bundle's value may be greater than $x + y$ owing to value-added by the bundling process. The bundle may be less valuable than the separate units if it contains an unwanted unit.

The larger is an individual's valuation of X when consumed alone, or of Y when consumed alone, the larger is the individual's valuation of the bundle B, that is, $b_x > 0$ and $b_y > 0$. When separately purchased, X and Y have a combined value equal to the sum of their stand-alone values x and y. Each individual's utility function is characterized by a combination of x, y, and b. We assume that x and y are distributed in the population

*AT&T-Bell Laboratories, Crawfords Corner Road, Holmdel, NJ 07733, and Duke University, Durham, NC 27706, respectively.

according to the density $f(x, y)$. We assume that $b(0, y) = y$ and $b(x,0) = x$.

Let $D(x, y) = (x + y) - b(x, y)$, then $D > 0$ if b is subadditive and $D < 0$ if b is superadditive. Let $\mathbf{D}$ be the set of all individuals whose utility for the bundle B exhibits additivity of degree less than $(P_x + P_y) - P_b$, that is,

$$(1) \quad \mathbf{D} = \{(x, y): D(x, y) \leq (P_x + P_y) - P_b\}.$$

The complement of $\mathbf{D}$ is denoted $\mathbf{D}^c$.

Depending on the commodity bundling strategy used by the firm, consumers will have a number of consumption options. In general, the options will include: 1) consume X alone and achieve a net utility of $x - P_x$; 2) consume Y alone and achieve a net utility of $y - P_y$; 3) consume the bundle B and achieve a net utility of $b(x, y) - P_b$; 4) separately purchase and consume both X and Y to achieve a net utility $(x + y) - (P_x + P_y)$; or 5) purchase nothing.

A. *Pure Bundling*

Under a pure bundling strategy, consumers may consume the bundle or nothing, that is, they choose between options 3 and 5. Define $y^b(x, P_b)$ such that $b(x, y^b) = P_b$, then consumers who prefer the bundle have reservation prices in the set

$$(2) \quad \mathbf{B}^0 = \{(x, y): y \geq y^b(x, P_b)\},$$

where $y^b(0, P_b) = P_b$. Consumers who purchase nothing have reservation prices in the complement of $\mathbf{B}^0$.

B. *Mixed Bundling*

Under mixed bundling, consumers will choose option Oj if the net utility associated with Oj is greater than the utility resulting from selection of any other option $Ok, k \neq j$. Thus X alone, option 1, is preferred if $x - P_x$ is greater than $y - P_y$, $b(x, y) - P_b$, and $(x + y) - (P_x + P_y)$, and $x - P_x$ is itself positive; these conditions respectively imply that option 1 is preferred to options 2, 3, 4, and 5. Similarly Y alone, option 2, is preferred if $y - P_y$ is greater than $x - P_x$, $b(x, y) - P_b$, $(x + y) - (P_x + P_y)$ and zero. The bundle, option 3, is preferred if $b(x, y) - P_b$ is greater than $x - P_x$, $y - P_y$, $(x + y) - (P_x + P_y)$, and zero. Self-bundling, option 4, is preferred if $(x + y) - (P_x + P_y)$ is greater than $x - P_x$, $y - P_y$, $b(x, y) - P_b$ and zero. Finally, the consumer will prefer option 5, that is, purchase nothing, if the net utility derived from all other options is negative.

Consumers who are indifferent to consuming X or the bundle B, have reservation prices $(x, y^e(x, P_b - P_x))$ where $b(x, y^e) = x + (P_b - P_x)$ for all x, with $y^e(0, P_b - P_x) = P_b - P_x$. Thus, consumers who would self-select X have reservation prices in the set

$$(3) \quad \mathbf{X} = \{(x, y): x > P_x, \; y < \operatorname{Min}[y^e(x, P_b - P_x), P_y]\}.$$

The bundle B is strictly preferred to X alone if (x, y) is such that $y > y^e(x, P_b - P_x)$, while x is preferred to B if $y < y^e(x, P_b - P_x)$. Indifference between Y and the bundle B obtains for reservation prices $(x^e(y, P_b - P_y), y)$ where $x^e(y, P_b - P_y)$ is defined by $b(x^e, y) = y + (P_b - P_y)$ with $x^e(0, P_b - P_y) = P_b - P_y$. Thus Y is preferred to B if $x < x^e(y, P_b - P_y)$; B is preferred to y if $x > x^e(y, P_b - P_y)$. The graph of reservation prices in $\mathbf{Y}$ is very similar to the graphs of $\mathbf{X}$. Hence consumers who choose to consume Y alone, under a mixed bundling strategy, will have reservation prices in

$$(4) \quad \mathbf{Y} = \{(x, y): x < \operatorname{Min}[x^e(y, P_b - P_y), P_x]\; y > P_y\}.$$

An individual who strictly prefers the bundle B to the separately purchased combination XY has reservation prices $(x, y) \in \mathbf{D}$. The combination XY is preferred to the bundle B by those individuals whose reservation prices (x, y) are in $\mathbf{D}^c$. It should be clear from the discussion of X and Y that an individual will choose to purchase *both* X and Y only if $x > P_x$ and $y > P_y$. Thus individuals who purchase both X and Y must have reservation prices in the set

$$(5)\quad \mathbf{XY}=\{(x,y): x>P_x, y>P_y\}\cap\mathbf{D}^c.$$

Therefore, if $P_b < P_x + P_y$ then **XY** is nonempty only if there exist (x, y) such that $D(x, y) > 0$ that is, there is some individual with subadditive preferences. If $P_b > P_x + P_y$ then **XY** is chosen by some individuals with superadditive preferences. It follows that if all individuals have additive preferences and $P_b < P_x + P_y$, as in previous literature, then **XY** is empty, that is, no individual would purchase both X and Y separately.

The set of consumers who choose the bundle, option 3, under a mixed strategy must have reservation prices in the set

$$(6)\quad \mathbf{B}=\{(x,y): x>x^e(y, P_b-P_y),$$
$$y>\mathrm{Max}[y^e(x,P_b-P_x), y^b(x,P_b)]\}\cap\mathbf{D}.$$

If $D(x, y) = 0\ \forall\ (x, y)$ and $P_b < P_x + P_y$ then **D** equals the positive quadrant, **B** is nonempty, and **XY** is empty. If $D(x, y) = 0\ \forall$ (x, y) and $P_b > P_x + P_y$, then both **B** and **D** are empty and $\mathbf{XY} = \{(x, y):\ x > P_x, y > P_y\}$. Hence if $b(x, y) = x + y$ for all individuals in the population, then for any given vector of prices (P_x, P_y, P_b), the firm cannot simultaneously have demand for both B and XY. This possibility does arise however, if there is diversity in consumers' valuation of the bundle relative to their valuation of its components. For example, Figure 1 shows a case in which there are positive demands for both B and XY, given that $P_b > P_x + P_y$.

In Figure 1, $D(x, y) = P_x + P_y - P_b$ marks the locus of indifference between B *and* XY, and dominates the influence of $y^b(x, P_b)$ and $x^e(y, P_b - P_y)$ on the specification of **B**. The locus of indifference between X and B, that is, $y^e(X, P_b - P_x)$, matters in the determination of **B**, when it falls below P_y. Consumers with reservation prices (x, y) in the dotted area choose the bundle while those in the diagonally shaded area choose to self-bundle. Consumers who respectively buy X or Y alone have reservation prices in the vertically or horizontally shaded areas.

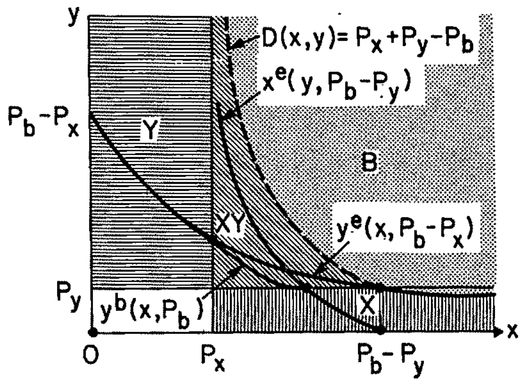


FIGURE 1. DEMANDS UNDER MIXED BUNDLE STRATEGY

II. Policy-Related Insights

The aggregate demand and profit functions, resulting from consumer self-selection, in this more general bundling model, are useful in addressing many of the policy issues mentioned in the introduction. We briefly discuss insights relating to two of these policy issues.

Suppose X and Y are nondurable, repeat purchase products. The firm pursues a pure bundle strategy, setting the bundle's price at P_b^0. A court or antitrust agency intervenes and prohibits use of the pure bundle strategy. The firm resorts to a legal, mixed bundle strategy with prices (P_x, P_y, P_b). Are the consumers of the bundle under the pure strategy, those in $\mathbf{B}^0(P_b^0)$, made better off by such intervention? Figure 2 shows that if $b(x, y)$ is subadditive of sufficiently high degree, then for any P_b^0 the firm can always select mixed bundle prices (P_x, P_y, P_b) such that almost all consumers in $\mathbf{B}^0$, the dotted area of Figure 2, continue to purchase the bundle under the mixed strategy.

Hence, if b is sufficiently subadditive, the prices can be chosen such that the number of consumers in the set

$$(7)\quad \Delta=\{(x,y): x>P_b^0-P_b+P_x, y<y^e\}$$
$$\cup\{(x,y):\ x<x^e, y>P_b^0-P_b+P_y\}$$

is arbitrarily close to zero, and $D(x,y^b(x,P_b^0))$

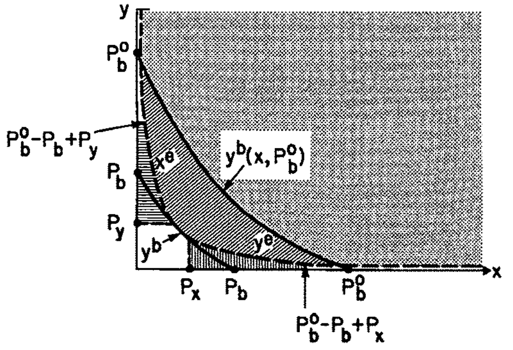


FIGURE 2. A RESPONSE TO BUNDLING PROHIBITION

$< P_x + P_y - P_b$ so that no consumer in $\mathbf{B}^0$ would choose to self-bundle under the mixed strategy. Yet, the firm will have positive demands for both X and Y, so that the mixed strategy does not degenerate to a pure strategy. Consequently, the firm is able to choose the mixed bundle prices so that consumers in $\mathbf{B}^0$ continue to purchase the bundle under the mixed strategy, even though the gross utility of the bundle is substantially less than the gross utility of unbundled consumption. Consumers in $\mathbf{B}^0$ may get a price break, however, if $P_b < P_b^0$, but this would not necessarily happen. In any case, consumers in $\mathbf{B}^0$ will not necessarily be made better off by prohibition of the pure bundle strategy unless restraints are also imposed on the firm's choice of prices. This raises a number of policy issues. Under what conditions will it be necessary to impose price restraints in order to make a bundling prohibition effective? What forms of price restraints, if any, best serve the public interest? When will pure bundle strategies be in the public interest? These questions can be dealt with using the model framework set forth in Section I. Indications are that antibundling enforcement is usually more effective when price restraints are also imposed. When there is sufficient diversity in consumers' valuation of the bundle, relative to its component parts, market segmentation by use of a bundle options mechanism can serve the interests of both the firm and consumers.

A second policy issue concerns the profits adjustments resulting from antibundling enforcement. If prohibitions against pure bundling have a relatively minor impact on profits but high intervention cost, then such antitrust activity may not serve the public interest. Adams-Yellen use an "Exclusion Principle" to establish whether additional profits can be made under a mixed bundle strategy when pure bundling is prohibited. A generalized version of this principle (no consumer buys the bundle when the increment to the value of the bundle contributed by one good exceeds the marginal cost of producing that good) can be applied to our model.

The profits impact of a switch from a pure to mixed bundle strategy can be seen as follows: let $N(\mathbf{S})$ denote the integral of $f(x, y)$ over the set $\mathbf{S}$, that is, $N(\mathbf{S})$ is the number of consumers who have reservation prices (x, y) in the set $\mathbf{S}$. The constant marginal cost of X and Y are denoted respectively by C_x and C_y; hence the marginal cost of the bundle is $(C_x + C_y)$. Profits, Π^0, at price P_b^0, under a pure bundle strategy, and profits Π^m, at prices (P_x, P_y, P_b), under a mixed bundle strategy, are, respectively,

$$(8)\quad \Pi^0 = \left[P_b^0 - (C_x + C_y)\right] N(\mathbf{B}^0)$$

$$\begin{aligned}\Pi^m = {} & [P_x - C_x][N(\mathbf{X}) + N(\mathbf{XY})] \\ & + \left[P_y - C_y\right][N(\mathbf{Y}) + N(\mathbf{XY})] \\ & + \left[P_b - (C_x + C_y)\right] N(\mathbf{B}).\end{aligned}$$

Consequently, the profits from a mixed bundle strategy are greater than profits from a pure bundle strategy if the mixed bundle's component prices at least equal their marginal cost, that is, $P_x \geq C_x$ and $P_y \geq C_y$, and if

$$(9)\qquad \frac{P_b - (C_x + C_y)}{P_b^0 - (C_x + C_y)} \geq \frac{N(\mathbf{B}^0)}{N(\mathbf{B})}.$$

Hence, if the ratio of the deviation of the bundle prices from their marginal cost is greater than the ratio of bundle demands, $N(\mathbf{B}^0)/N(\mathbf{B})$, then Π^m is greater than Π^0.

The profitability of bundling vs. unbundled sales is then shown to depend on the degree of additivity of $b(x, y)$ as well as the distribution $f(x, y)$ of reservation prices for X and Y. We have constructed examples in which reservation prices are perfectly, positively correlated, yet bundling is the preferred strategy if $b(x, y)$ is superadditive. If $b(x, y)$ is subadditive, then mixed bundling appears more profitable than pure bundling, whenever there are consumers, (who buy the bundle) for whom the degree of additivity $D(x, y)$ is greater than the marginal cost of producing the bundle.

This analysis suggests that the evaluation of the legality of tie-in sales should include an assessment of the incremental value of the bundle. To some extent, courts have done this, by recognizing that tie-ins may improve or maintain product quality. A broader issue is the impact of the prohibition of tie-in sales on prices, profits and economic efficiency. Since pure commodity bundling is prohibited, but mixed bundling is not, it is possible that a firm can negate the impact of the prohibition, unless price restraints are also imposed.

REFERENCES

Adams, Williams J. and Yellen, Janet L., "Commodity Bundling and the Burden of Monopoly," *Quarterly Journal of Economics*, August 1976, *90*, 475–98.

Craswell, Richard, "Tying Requirements in Competitive Markets: The Consumer Protection Issues," *Boston University Law Review*, May 1982, *62*, 661–700.

Lewis, David A., "Whether Patented or Unatented: A Question of the Economic Leverage of Patents to Coerce Tie-Ins," *Journal of Law and Technology*, March 1982, *23*, 77–106.

Palfrey, Thomas R., "Bundling Decisions by a Multiproduct Monopolist with Incomplete Information," *Econometrica*, March 1983, *51*, 463–84.

Ross, Michael E., "The Single Product Issue in Antitrust Tying: A Functional Approach," *Emory Law Journal*, April 1982, *25*, 67–71.

Schmalensee, Richard A., "Commodity Bundling by Single Product Monopolies," *Journal of Law and Economics*, April 1982, *25*, 67–72.

Stigler, George J., "A Note on Block Booking," in *The Organization of Industry*, Homewood: R. D. Irwin, 1968.

Lessons from the 1979–82 Monetary Policy Experiment

By BENJAMIN M. FRIEDMAN*

Macroeconomics is not a laboratory science. Economists must learn about macroeconomic behavior from the events that occur in the real world, rather than from controlled experiments that they can design and implement themselves. Especially when events represent potentially substantial breaks from prior experience—in other words, when they greatly increase the range and variance of the available data—such real world "experiments" can provide important information about how economies, and the households and firms that comprise them, behave. The quadrupling of oil prices in the early 1970's was one example, and economists have been quick to learn from it. The experience of U.S. monetary policy at the outset of the 1980's has now provided another such opportunity.

The latest monetary policy experiment in the United States lasted almost precisely three years. On October 6, 1979, the Federal Reserve System announced a new policy orientation in which it would henceforth place renewed emphasis on growth targets for the major monetary aggregates, and also implement new operating procedures to help achieve those targets. The principal motivation for these changes was an economic situation marked by rising price inflation, already at or near record postwar levels, and a deteriorating international value of the dollar. On October 9, 1982, the Federal Reserve chairman announced a "temporary" abandonment of the stated growth target for the narrow *M*1 money stock, up to then by far the most important monetary aggregate for policy purposes. The economic situation motivating this reversal was a deepening business recession with unemployment at record levels, despite money growth in excess of targeted ranges.

The object of this paper is to survey the lessons to be drawn from this three-year monetary policy experiment. The focus is on lessons associated with the overall use of monetary aggregate targets, rather than the specific operating procedures used to achieve them. The plan of the paper is to consider a series of familiar propositions often (but certainly not universally) associated with the use of monetary aggregate targets for monetary policy, in light of various forms of evidence from these three years ranging as seems appropriate (given space limitations) from simple inspection of data series to more elaborate statistical procedures. To anticipate, the evidence from the 1979–82 experiment leads to doubt, rather than confidence, in each of these propositions.

I. Money and Nominal Income

To begin, targeting monetary aggregates requires deciding what monetary aggregates to target. In the short run—say, a year or so—shifts in the portfolio preferences of the nonbank public may change the relationships defining mutually consistent growth rates for different deposit-type aggregates. Over longer time horizons, however, like those relevant for "gradualist" proposals to slow money growth by, say, 1 percent per annum until price inflation is eliminated, it would be convenient for policy purposes to believe

PROPOSITION 1: *The major monetary aggregates move roughly together over substan-*

*Harvard University, Cambridge, MA 02138. I am grateful to Diane Coyle and Kenneth Weiller for research assistance, and to the National Science Foundation and the Alfred P. Sloan Foundation for research support.

TABLE 1—GROWTH OF MONEY AND NOMINAL INCOME, 1978–82[a]

	$M1$	$M2$	$M3$	Nominal Income
1978	8.1	8.0	11.3	14.7
1979	7.4	8.1	9.6	9.7
1980	7.2	9.0	9.7	9.3
1981	5.1	9.4	11.7	10.8
1982	8.5	9.3	10.1	2.6

[a]Shown in percent.

tial spans of time, so that the central bank can simply pick one aggregate and target it appropriately without having to worry about mixed signals.

Table 1 shows the annual growth rates (fourth quarter over previous fourth quarter, as the Federal Reserve formulates its targets) for the three major monetary aggregates during 1978–82. Even over a half-decade, the basic directions indicated respectively by $M1$, $M2$, and $M3$ disagreed. Given the inherited history of 1978, the Federal Reserve under the new policy did approximately achieve a 1 percent per annum slowing in $M1$ growth over 1979, 1980, and 1981. By contrast, $M2$ growth became consistently faster during these years, while $M3$ growth fluctuated without discernable trend. Even in 1982, the year in which the experiment ended, the widely discussed easing of monetary policy is apparent in a quickening of $M1$ growth but not $M2$ growth, while $M3$ growth slowed sharply.

The point here is not to determine whether these dissimilar growth rates are explainable in terms of accepted portfolio-theoretic behavior in conjunction with the financial innovations and regulatory changes affecting the U.S. banking system during these years. It is instead that, even over a five-year period, the answer to so basic a question as whether money growth is speeding up or slowing down depends on which among the major monetary aggregates is doing the answering. The implication for monetary policy is that "monetary aggregates" (i.e., the major aggregates *collectively*) are of limited usefulness as a central focus for policy. To use monetary aggregates in this way, the central bank must have a clear view of which specific aggregates it is using, and why.

At a more basic level, placing monetary aggregates at the center of the monetary policy process depends not just on their relationships among themselves, but on their connection to nonfinancial economic activity. An important line of thinking has argued that it is appropriate to think about this connection, at least at the outset, in terms of a relationship between money and nominal income. Once again, for most policy purposes it is not important—or, given feasible monetary control, even very relevant—to have a tight relationship over short time periods. Nevertheless, for time horizons like those involved in the recent experiment, it is difficult to motivate the use of monetary aggregate targets for monetary policy without claiming

PROPOSITION 2: *The movement of at least some monetary aggregate roughly explains the movement of nominal income over substantial time spans.*

Table 1 also shows the annual growth (again, fourth quarter over previous fourth quarter) of nominal Gross National Product. Even after making allowance for plausible time lags, it is difficult to examine the data in Table 1 as a whole and conclude that the movement of any one monetary aggregate has even roughly accounted for the movement of nominal income over these years. The best candidate for explaining the 5 percent fall in income growth in 1979 is $M2$ growth, which declined from 11.2 percent in 1977 (not shown) to 8.0 percent in 1978.[1] By contrast, $M1$ growth is the only one of the three to have declined in 1981, and even that decline is small in comparison with the more than 8 percent fall in income growth in 1982.

In sum, the movement of nominal income during these years has been difficult to reconcile with the respective movements in the major monetary aggregates, at least without going well beyond the usual arguments for

[1]Neither $M1$ growth nor $M3$ growth showed much slowing in 1978.

monetary aggregate targets based on the presumption of a stable (and, usually, a relatively interest insensitive) money-income relationship.

II. Price Inflation and Real Economic Growth

Presumably policymakers care not just about nominal income growth but also about price inflation and real growth separately. One of the most interesting developments in macroeconomics within the past decade has been a line of reasoning implying that, because of effects due to expectations, the use of pre-announced monetary aggregate targets may favorably affect the respective impacts of monetary policy on inflation and real economic activity. In the context of a disinflation through monetary policy like that begun in the United States in 1979, the idea is that a slowing of monetary growth that is widely publicized in advance, as in October 1979 and thereafter, would affect the expectations on which households and firms act, and thereby cause a given slowing of nominal income growth to consist of more rapid slowing of inflation, and less slowing of real activity, than would otherwise be the case.

If valid, this role of monetary aggregate targets would be valuable indeed. Just as the idea of the stable Phillips curve held out the prospect of solving the chief macroeconomic policy problem of the 1950's and 1960's, unemployment, without the cost of accelerating inflation, the "new classical macroeconomics" has more recently offered the prospect of solving the chief macroeconomic problem of the 1970's, inflation, without the conventionally associated costs of foregone output, employment, and income.

This view of the potential contribution of pre-announced monetary aggregate targets involves several elements on which the 1979–82 experiment in U.S. monetary policy can shed light. One, following familiar criticisms of the standard Phillips curve literature, is

PROPOSITION 3: *A pre-announced slowing of money growth will lead to a more rapid slowing of price inflation than would be consistent with prior historical correlations.*

The actual path of U.S. price inflation since October 1979, in comparison to forecasts from equations based on prior data, shows just the opposite. The small "structural" macroeconometric model estimated using quarterly 1961:1–1979:3 data in Richard Clarida's and my article (1983) includes a simple linear function relating price inflation to lagged values of real growth, changes in the terms of trade, and inflation itself.[2] Although the relevant *F*-test for the null hypothesis of stable coefficients provides marginally significant evidence of a break with the onset of the new monetary policy regime in 1979:4, the equation's dynamic forecast for 1979:4–1983:2 indicates that this break has been *in the opposite direction* to that implied by the new classical macroeconomics. The equation *under*predicts inflation in fourteen of the fifteen forecast quarters, with an overall average predicted inflation rate of only 5.0 percent per annum vs. the actual 7.0 percent. The slowing of price inflation since October 1979 has been not more rapid but more sluggish than would have been consistent with the correlations exhibited by prior experience, given the subsequent two business recessions and the sharp appreciation of the exchange rate. In other words, what has been surprising about inflation during this period was how sluggishly, not how rapidly, it slowed. Similar price equations (see, for example, George Perry, 1983) show similar results.

Moreover, this result is not simply due to an arbitrarily specified set of "structural" restrictions on the data. The vector autoregression model estimated using data through 1979:3 in Clarida's and my article (1984) includes the inflation rate, the respective growth rates of real income, money (*M*1) and total credit, and the changes in the

[2] The equation is

$$\Delta P_t = \underset{(3.4)}{.0895}\Delta X_{t+1} + \underset{(3.9)}{.0542}\Delta I_{t-1} + \underset{(25.2)}{.8700}\Delta P_{t-1},$$

$$SE = .00347,\ \overline{R}^2 = .88,\ \rho = -.1,$$

where P is the *GNP* deflator, X is real *GNP*, I is the dollar price of imports, and all variables are in natural logarithms.

Treasury bill rate and the federal deficit. The unconditional dynamic forecast generated by this completely nonstructural way of summarizing the correlations in the pre-1979:4 data overpredicts inflation (8.4 percent per annum) on average during 1979:4–1983:2, but the forecasting exercise which most closely corresponds to the proposition relating the slowing of inflation to the use of pre-announced monetary targets does the opposite. In particular, using a technique due to Thomas Doan et al. (1983) to forecast inflation during each quarter of 1979:4–1983:2 on the basis of the historical correlations as summarized by the model as well as the actual values of money growth in all quarters of this period raises the mean forecast by an absurd amount (to 24 percent) in comparison to either the actual experience or the unconditional forecast.[3]

The other side of the coin of favoring the use of pre-announced monetary aggregate targets because the associated expectations effects may make disinflation more rapid is that they may make it less costly. Conventional estimates, like those summarized by Arthur Okun (1978), have indicated that the cost of each one percentage-point reduction in the ongoing rate of price inflation achieved via monetary policy is between two and six "point-years" of unemployment, with a median estimate of three point-years (or, equivalently, 6–18 percent of a year's total output, with a median of 9 percent). Such pessimistic estimates have often discouraged advocates of disinflationary monetary policy. By constrast, the same reasoning associating a pre-announced slowing of money growth with unexpectedly rapid disinflation suggests

TABLE 2—PERCENTAGE RATES OF INFLATION, GROWTH, AND UNEMPLOYMENT, 1978–83

	Price Inflation	Real Growth	Unemployment Rate	Cumulative Excess Unemployment
1978	7.4	5.0	6.1	–
1979	8.6	2.8	5.8	–
1980	9.3	−0.4	7.1	1.1
1981	9.4	2.6	7.6	2.7
1982	6.0	−1.9	9.7	6.4
1983	5.0[a]	3.4[a]	9.7[b]	10.1

[a] First three quarters at annual rate.
[b] First eleven months.

PROPOSITION 4: *A pre-announced slowing of money growth will cause a given slowing of price inflation to be accompanied by less foregone output, employment, and income than would be consistent with prior historical correlations.*

Table 2 shows the annual rates of change of real Gross National Product and the associated price deflator, and the annual average unemployment rate during 1978–83. The final column of the table also shows, for 1980–83, the cumulative excess of the unemployment rate above 6 percent (the approximate average for the two prior years). The slowing of inflation from near 10 percent in 1980–81 to 5 percent in 1982 has, *just during 1980–83*, required some 10 point-years of excess unemployment—about at the 2-to-1 lower end of the range surveyed by Okun. Stopping the accounts at 1983 makes no sense, however. Even the optimistic view that the U.S. economy will return to full employment fairly quickly, with no reversal at all in the disinflation already achieved, puts the likely final tally closer to Okun's 3-to-1 median. If the current economic recovery falters, or if inflation speeds up, the final tally could easily be nearer the 6-to-1 upper end.

Whether this ratio ultimately turns out to be somewhat above or somewhat below Okun's median is beside the point. What matters is that the real costs of disinflation achieved by monetary policy have been about in line with earlier conventional estimates,

[3] A crucial question, of course here and below), is whether households and firms believed that monetary policy would take the course it did 1979–82. Perhaps the best that can be said is that, if the experiment of these years was not an example of the kind of "regime change" to which new classical macroeconomics arguments are supposed to be relevant, then it is not clear to what real world event they ever would be relevant. Thomas Sargent and Neil Wallace 1981) have made a potentially important qualification to the usual results as stated above, noting the necessity of a consistent accompanying fiscal policy; but the federal government's budget on a high-employment basis showed only small deficits in 1980 and 1982, and a surplus in 1981.

notwithstanding the use of monetary aggregate targets.

III. Monetary Policy and Long-Term Interest Rates

Price inflation and real growth are not the only dimensions of economic activity for which the impact of monetary policy depends importantly on expectations. Perhaps the most familiar aspect of this subject is the behavior of the yields on (prices of) assets which represent explicit future claims, and which therefore explicitly involve expectations about future events. A standard distinction in this context is that between the respective effects of monetary policy on short- and long-term interest rates. While a tightening of monetary policy might well lead to higher short-term rates (unless the new classical macroeconomics arguments discussed above are valid), it need not lead to higher long-term rates if those rates embody expectations of lower price inflation (and hence lower short-term rates) in the future. More specifically, the idea is

PROPOSITION 5: *A pre-announced slowing of money growth will lead to lower long-term interest rates than would ordinarily be consistent with the prevailing levels of short-term rates, given prior historical correlations.*

The actual path of U.S. long-term interest rates since October 1979, in comparison to forecasts from equations based on prior data, shows the opposite. The "structural" model estimated using 1961:1–1979:3 data in Clarida's and my article (1983) includes a simple linear function relating the bond rate to current and lagged values of the Treasury bill rate, lagged changes in the maturity composition of outstanding federal government debt, and the lagged bond rate.[4] As with the model's price equation, there is significant evidence of a break after 1979:3, but here too the observed shift has been *in the opposite direction* to that implied by the proposition about the use of monetary aggregate targets. The equation *under*predicts the bond rate in every quarter during 1979:4–1983:2, with an overall average predicted rate of only 10.91 percent vs. the actual 14.81 percent. Given short-term rates, long-term rates have been surprisingly high, not surprisingly low. Other term structure equations (see, for example, Robert Shiller et al., 1983) show similar results.

Finally, ever since the Federal Reserve began to focus great attention on its monetary aggregate targets, a familiar argument has been that market participants' expectations have rendered the central bank a "prisoner" to its own announcements. The basic reasoning involved has been just the inverse of that examined above, again denying the ability of monetary policy to affect long-term interest rates except by affecting expectations of future price inflation. Any easing of monetary policy involving money growth significantly in excess of the targeted range would lead long-term interest rates to rise rather than fall. In the extreme case, the expectation associated with the abandonment of such targets is

PROPOSITION 6: *Abandonment of monetary aggregate targets for monetary policy, especially in conjunction with money growth in excess of previously targeted ranges, will cause long-term interest rates to rise.*

The movement of U.S. long-term interest rates that accompanied the end of 1979–82 monetary policy experiment was just the opposite. The Federal Reserve began its move toward a degree of ease not warranted by the money growth targets shortly after midyear 1982, and on October 9 the chairman publicly

[4] The equation is

$$\begin{aligned} r_{Lt} = \underset{(1.4)}{.0472} + \underset{(1.1)}{.1441} r_{St} - \underset{(-0.5)}{.0579} r_{S,t-1} + \underset{(2.3)}{.1376} (L-S)_t \\ + \underset{(37.0)}{.9100} r_{L,t-1} \end{aligned}$$

$$SE = .020, \overline{R}^2 = .98, \rho = .4,$$

where r_L is the Baa bond rate, r_S is the 3-month Treasury bill rate, L and S are the respective amounts of long-and short-maturity federal government debt outstanding, and all variables are in natural logarithms. The coefficients on current and lagged r_S are highly significant jointly, though not individually.

announced the "temporary" abandonment of the $M1$ target. The Baa bond yield declined from 16.78 percent in 1982:2 to 16.25 percent in 1982:3, as market participants began to infer that policy had changed, and the further decline to 14.39 percent in 1982:4 constituted the largest one-quarter rally in the postwar experience of the U.S. fixed-income securities market. The decline continued further, to 13.25 percent in 1983:2, as money growth became still faster.

Participants in the U.S. securities markets are apparently more sensible than to hold monetary policy prisoner to a counterproductive policy structure. When the Federal Reserve abandons a policy that is not working, the market records its approval.

REFERENCES

Clarida, Richard H. and Friedman, Benjamin M., "Why Have Short-Term Interest Rates Been So High?," *Brookings Papers on Economic Activity*, 2:1983, 553–78.

______ and ______, "The Behavior of U.S. Short-Term Interest Rates since October 1979," *Journal of Finance*, forthcoming 1984.

Doan, Thomas, Litterman, Robert and Sims, Christopher A., "Forecasting and Conditional Projection Using Realistic Prior Distributions," mimeo., National Bureau of Economic Research, 1983.

Okun, Arthur M. "Efficient Disinflationary Policies." *American Economic Review Proceedings*, May 1978, *68*, 348–52.

Perry, George L. "What Have We Learned About Disinflation?," *Brookings Papers on Economic Activity*, 2:1983, 587–602.

Sargent, Thomas J., and Wallace, Neil, "Some Unpleasant Monetarist Arithmetic," *Federal Reserve Bank of Minneapolis, Quarterly Review*, Fall, 1981, *5*, 1–17.

Shiller, Robert J., Campbell, John Y. and Schoenholtz, Kermit, "Forward Rates and Future Policy: Interpreting the Term Structure of Interest Rates," *Brookings Papers on Economic Activity*, 1:1983, 173–217.

Monetarist Rules in the Light of Recent Experience

By BENNETT T. MCCALLUM*

The title of this session, like a host of recent writings by critics of monetarism, suggests that the period from late 1979 to mid-1982 witnessed a significant "monetarist experiment," that is, that U.S. monetary policy during that period conformed to monetarist prescriptions. For a number of reasons, however, that suggestion is clearly untenable. Among these are the following: growth rates of monetary aggregates fluctuated widely on a month-to-month or quarter-to-quarter basis; the Fed's operating procedures were more poorly designed for money stock control than those in place prior to October, 1979;[1] and discretionary responses to current cyclical conditions were never foresworn. In addition, the growth rate of the ($M1$) money stock was only slightly lower than that of the previous decade and was higher than that of the previous twenty years.[2] It is true that the Fed demonstrated considerable resolve in reducing $M1$ growth rates from the values of 1977–78, even though this required a spell of unusually high interest rates, and the operating procedures announced in October 1979 were politically helpful in disclaiming responsibility for these unpopular rates. But such steps hardly constitute an embrace of monetarism—an opinion shared, it might be added, by officials of the Fed as well as leading proponents of monetarism.[3]

Consequently, any argument of the form "we tried monetarism and it produced undesirable results" seems to me unworthy of discussion. But that conclusion does not eliminate the possibility that the time period since 1979 has produced new evidence relevant to a reasoned consideration of the desirability of monetarist prescriptions. It is, in other words, conceivable that the new data points generated during that period were so highly informative about the nature of the economy that opinions concerning the merits of monetarist prescriptions could reasonably have been altered. Let us then continue the discussion from that perspective.

I. New Evidence

Contemplating the developments of 1979–82, one is led to the conclusion that the main new information relevant to monetarism does not pertain to the behavior of macroeconomic variables or their interrelationships. Much of the discussion has, admittedly, emphasized the unusually high interest rates (nominal and *ex post* real) of 1980–81, the severe unemployment of 1982, the rapid decline in inflation between 1981 and 1982, and the sharp fall in $M1$ velocity during 1982. But none of these facts is incompatible with the basic hypotheses about the economy that are essential to a monetarist position, namely, that money stock movements have strong effects on nominal GNP and that there is no permanent tradeoff between unemployment and inflation.[4] Nor do these facts serve

*Professor of Economics, Carnegie-Mellon University, Pittsburgh, PA 15213; Research Associate, National Bureau of Economic Research; and Adviser, Federal Reserve Bank of Richmond. I am indebted to Marvin Goodfriend and Allan Meltzer for helpful suggestions, and to the National Science Foundation (SES 82-08151) for financial support.

[1] This is suggested by the analytical results in my article with James Hoehn (1983), as well as the discussion in Karl Brunner and Allan Meltzer (1983), and elsewhere.

[2] From December 1979 through June 1982, the average $M1$ growth rate was 6.1 percent. During 1979–79, the figure was 6.6 percent; for 1960–79, it was 5.2 percent. (Data from *Economic Report of the President*, February 1983.)

[3] See, for example, contributions by Paul Volcker, Milton Friedman, and Meltzer in the Joint Economic Committee's *Monetarism and the Federal Reserve's Conduct of Monetary Policy* (1982).

[4] These hypotheses are discussed in my 1981 paper. That neither the second or third of the four cited facts is inconsistent with monetarist principles should not need to be argued. The first fact would be inconsistent with the "Ricardian" brand of monetarism *if* it were established that the high interest rates of 1980–81 resulted from unmonetized deficits, but such is only one hypothesis. Also, less extreme forms of monetarism are

to discredit—indeed, quite the contrary!—the monetarist presumption that no one possesses a reliable structural model of the economy. Instead, the new developments that deserve attention are those pertaining to technological innovation and regulatory change in the payments industry, prominent examples of which include the introduction of NOW and sweep accounts, the growth of money market mutual funds, and the widespread use of repurchase agreements. These developments have, as is well-known, led to redefinitions of traditional monetary aggregates such as $M1$ and $M2$, and have weakened the correspondence of the former to the concept of the medium of exchange. As a consequence, the developments have engendered a widespread belief that the economy's demand function for any operationally defined monetary aggregate has been subject to significant shifts and will continue to shift in an unpredictable fashion in the future.

In saying this, I do not mean to deny that these developments have in large part been a response to prevailing regulations and policy. Nor do I mean to express agreement with antimonetarist observers who justify each departure from announced monetary targets with a claim that money demand has shifted, or those who believe that all tangible media of exchange will soon disappear as modern economies adopt accounting systems of exchange. Nor do I wish to suggest that technical progress in the payments industry is something new, which it surely is not. But I nevertheless believe that recent experiences have served to reinforce pre-existing reasons for doubting that the best way of expressing monetarist prescriptions is in the form of a constant growth rate rule for the money stock. While the latter—whether measured as $M1$ or $M2$—could certainly be made more controllable by the Fed, wide-ranging institutional changes such as those recommended in Milton Friedman's *Program for Monetary Stability* (1960) would be required for truly tight control.[5] Consequently, it will probably continue to be the case that the money stock, which is not an ultimate *goal* variable, is also not a directly manipulable monetary *instrument*.

II. A Revised Monetarist Rule

Because of this intermediate status of the money stock—identifiable as neither instrument nor goal variable—its role is undesirably ambiguous. It is thus not surprising that money stock "targeting" as practiced by the Fed has been characterized by ambiguity, with departures from target values sometimes treated as something to be eliminated but often as mere bits of information requiring no response. It would be better, it would appear, to have a rule that specifies behavior of the monetary *base*, or stock of high-powered money, which is directly enough under Fed control to be treated as a bona fide instrument.[6] With observations on the base obtainable from Fed and Treasury balance sheets, its magnitudes could be monitored almost continuously and no significant departures from specified target paths would ever need to occur.

It will be objected that the velocity of the base relative to *GNP* will again be changing irregularly in the future so that a constant growth rate for the base would be undesirable. But there is nothing in the concept of a rule that requires the growth rate to be constant. Personally, I suspect that constant base growth at 1 percent per year over the next

more typical. And as for the fourth fact, a change in velocity does not imply a shift in money demand behavior—a sharp drop in interest rates (as in 1982) should induce a velocity decline according to most theories. Furthermore, evidence of behavioral shifts taken from studies of conventional money demand functions is unsatisfactory because estimates of the usual (Stephen Goldfeld, 1976) specification are implausible, implying as they do that many quarters are needed for portfolio adjustments that can be effected almost instantaneously with negligible cost.

[5] Furthermore, it would appear that extremely good substitutes for bank deposits can be developed by intermediaries not subject to any given set of regulations.

[6] The importance of controllability and the absence of ambiguity play crucial roles in Milton Friedman's argument for a rule expressed in terms of the money stock, rather than the price level (1960, pp. 86–89). Emphasis on the monetary base has long been recommended by Brunner and Meltzer. Another possible instrument is total reserves.

decade would yield satisfactory macroeconomic performance. But still better performance should be provided by a rule that periodically adjusts the base growth rate in response to past movements in some nominal target variable, with nominal *GNP* an attractive candidate. Thus in my opinion a desirable monetarist rule would adjust the base growth rate each month or quarter, increasing the rate if nominal *GNP* is below its target path, and vice versa. This target path, in turn, should specify nominal *GNP* growth of about 3 percent per year, a figure consistent with near-zero inflation.

There are three distinct ways, deserving of explicit mention, in which macroeconomic performance could benefit from an adjustable growth rate rule as compared with one of the constant-growth rate (*CGR*) type. First, the adjustable scheme could correct for any tendency of base velocity growth to change secularly as the pace of technological innovation increases or decreases. Second, there would be stronger countercyclical effects on aggregate demand and these would be of an automatic type.[7] Third, the adjustments would counteract the tendency (discussed in my earlier paper) of *CGR* rules plus fixed tax schedules to generate dynamic instability in the stock of government debt.

I am confident that both monetarists and antimonetarists will object to this proposal, partly for substantive reasons and partly because I have termed it "monetarist." Taking the last objection first, the label is warranted because the proposed rule is entirely nondiscretionary and is expressed exclusively in terms of nominal magnitudes. Substantively, the only apparent drawback from a monetarist perspective is that the adjustable base growth rate lacks the popular appeal of an absolutely constant value. While an adjustable path would be slightly harder to monitor than a constant path if everything else were the same, having a path expressed in terms of an instrument variable should provide greater operational content and improved monitoring possibilities in comparison with a constant growth path for a variable that can only be controlled indirectly.

The substantive objections from nonmonetarists will be different, of course, and will presumably focus on the nondiscretionary aspect of the proposal. In this regard, it is important to recognize that while the rule is nondiscretionary, it is not nonactivist. That is, the rule specifies instrument settings for the base growth rate that are contingent on the state of the economy—nominal *GNP* relative to its target path. It is less activist than a rule that (for example) prescribes higher-than-average nominal *GNP* growth when unemployment is above average, but it is activist nonetheless. An important reason for not attempting the more ambitious type of rule is the absence of a reliable model of the economy. If such a model were available, it would make no sense to use an intermediate target variable like nominal *GNP*, rather than focussing directly on ultimate goals (Benjamin Friedman, 1975).

III. Rules vs. Discretion

There remains to be discussed why the Fed should adopt any rule at all, rather than proceeding in a discretionary manner, as the latter would permit the same month-to-month base growth values as the rule would dictate, but would also provide the virtue of *flexibility*. The answer is that, given the nature of the U.S. economy, flexibility is not a virtue; the absence of flexibility could lead to superior performance in terms of unemployment-inflation combinations. To make this argument as simply as possible, I will utilize an analogy.[8] Imagine parents confronted with an instance of misbehavior by their child, whose welfare is uppermost in their minds.

[7] If the economy is in fact Keynesian in nature, these countercyclical effects would be likely to be helpful. If it is in fact purely classical, they should be neither helpful nor harmful.

[8] This analogy was suggested to me by Stanley Fischer. The argument that it is supposed to elucidate was originally developed by Finn Kydland and Edward Prescott (1977) and was usefully elaborated by Robert Barro and David Gordon (1983). The reader is referred to these papers for a more formal analysis, including a more precise specification of the assumed nature of the economy. The argument abstracts from cyclical fluctuations only for simplicity.

Shall they punish the child for the misbehavior, thereby inflicting disutility on themselves as well as the child? Or shall they refrain from punishment in this particular case, while promising punishment for all future instances of misbehavior? From the viewpoint of the moment in time when this decision is being faced, it is clearly optimal to select the second of these options. But of course the same choice will turn out to be optimal after the *next* instance of misbehavior, and so on, so the resulting steady state is one in which the parents never punish the child, whose behavior accordingly conforms to a regime in which there is no need ever to fear punishment.

In an economy with widespread nominal contracting, the problem faced each "year" is similar to that of parents confronted with an instance of misbehavior. The options are to impose monetary stringency, with resulting disutility for most parties, or to refrain from stringency this year while promising stringency in all future years. But with decision-making flexibility, the same choice will be made in future years when the intervening year's misbehavior becomes a thing of the past. Thus stringency tends to be imposed rarely, yet—since there is no permanent stimulus to employment from monetary leniency—there is no additional employment to compensate for the additional inflation that results from monetary leniency.

Some parents, however, obtain superior outcomes by sacrificing flexibility—by not making their choices after each instance of misbehavior, but instead adopting a rule that results in automatic punishment after each case of misbehavior.[9] The Fed could similarly obtain superior outcomes by surrendering flexibility in favor of a rule of the type described above. To do so would not only result in improved economic performance, but would also represent genuine *policy* behavior—as opposed to case-by-case attempts to optimize—of the type that central bank independence is intended to produce. Thus, despite the political pressures described by Edward Kane (1982) and others, the Fed should have a powerful motive to adopt a policy rule: if choices are to be made on a case-by-case basis, there is no reason why they should be made by an independent agency instead of the current administration.

[9] This is only a hypothetical possibility, of course; I do not claim to know of any actual cases.

REFERENCES

Barro, Robert J. and Gordon, David B., "A Positive Theory of Monetary Policy in a Natural-Rate Model," *Journal of Political Economy*, August 1983, *91*, 589–610.

Brunner, Karl and Meltzer, Allan H., "Strategies and Tactics for Monetary Control," *Carnegie-Rochester Conference Series on Public Policy: Money, Monetary Policy and Financial Institutions*, Spring 1983, *18*, 59–104.

Friedman, Benjamin M., "Targets, Instruments, and Indicators of Monetary Policy," *Journal of Monetary Economics*, October 1975, *1*, 443–73.

Friedman, Milton, *A Program for Monetary Stability*. New York: Fordham University Press, 1960.

Goldfeld, Stephen M., "The Case of the Missing Money," *Brookings Papers on Economic Activity*, 3:1976, 683–730.

Kane, Edward J., "Selecting Monetary Targets in a Changing Financial Environment," *Monetary Policy Issues in the 1980s*, Federal Reserve Bank of Kansas City, 1982.

Kydland, Finn E. and Prescott, Edward C., "Rules Rather than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, *85*, 473–91.

McCallum, Bennett T., "Monetarist Principles and the Money Stock Growth Rule," *American Economic Review Proceedings*," May 1981, *71*, 134–38.

______ **and Hoehn, James G.,** "Instrument Choice for Money Stock Control with Contemporaneous and Lagged Reserve Requirements," *Journal of Money, Credit, and Banking*, February 1983, *15*, 96–101.

U.S. Congress, Joint Economic Committee, *Monetarism and the Federal Reserve's Conduct of Monetary Policy*. Washington: USGPO, 1982.

Did Financial Innovation Hurt the Great Monetarist Experiment?

By JAMES L. PIERCE[*]

In October 1979, the Federal Reserve announced a new commitment to fight inflation. It signalled its resolve by a shift in policy tactics that placed greater emphasis on reducing the growth of money and less emphasis on limiting short-run fluctuations in interest rates. This shift in tactics was accomplished by a change in operating procedures that placed primary emphasis on controlling the growth of reserves available to depository institutions while greatly expanding the allowable range of fluctuations in the federal funds rate.

The growth of *M*1 and other monetary aggregates did slow on average, but money growth experienced large short-term fluctuations. Interest rate movements increased, as advertised, but the extent of their fluctuation was severe. Inflation slowed markedly, but this was primarily the consequence of a severe recession. Events gave grim testimony to the truth that it is possible to reduce inflation quickly by producing a sufficiently severe recession.

In the second half of 1982, with the economy in disarray and with growing concern about the ability of the financial system to withstand further strain, the Federal Reserve abandoned its new operating procedures. There was a shift back to the more "comfortable" world of stabilizing fluctuations in the federal funds rate.

There is considerable controversy over whether the Federal Reserve actually pursued a policy strategy that was consistent with the teachings of the "Book of Monetarism." Concern about limiting money growth and using reserves as the operating variable are consistent with monetarism. The extreme fluctuations in money growth during the period are not consistent with monetarist doctrine, however. Perhaps the Fed had not really embraced monetarism. It may have found that focusing on money growth was a convenient means of absolving itself from responsibility for the record-high interest rates that occurred. Conversely, perhaps the Fed did embrace the principles of monetarism, but was unable to achieve steady money growth. We shall never know for sure whether or not the Federal Reserve was *really* trying to perform a monetarist experiment on the American economy. We do known, however, that the experiment was far from pure because of the substantial money-growth volatility that occurred.

There is also substantial controversy over the success of the experiment. Some observers argue that the primary aim of reducing inflation was achieved; they proclaim the experiment a success. Others point out that this "success" was the consequence of a severe recession produced by high real interest rates and had nothing to do with monetarism per se.

This paper does not contribute to the debates concerning the nature and success of the "great monetarist experiment." Rather, it looks at the role played by financial innovation in complicating the pursuit of monetary policy during the period. Was the great experiment hurt by financial innovation? More specifically, did rapid financial innovation make it infeasible to target on *M*1 (or some other monetary aggregate), and did innovation contribute to the extreme fluctuations in money growth and interest rates that occurred? The conclusions from the discussion that follows is that financial innovation did not make it any less feasible to target on *M*1 during 1979–82 than for other periods. Financial innovation does not appear to have made money growth an unusually unreliable target, and innovation was not the source of the volatility of money growth and interest rates that occurred.

I

Before chronicling recent financial innovations and assessing their effects, it may be

[*] Professor of Economics, Department of Economics, University of California, Berkeley, California 97420.

useful to discuss briefly the general effects of financial innovation and to consider a timely example. The financial innovations that have occurred in recent years have lowered transactions costs for asset purchases and sales, they have introduced new assets and liabilities, and they have allowed depository institutions to avoid reserve requirements and other regulations. The innovations produce shifts in asset demands and liability supplies, and they alter elasticities of demand and supply. These changes in the structural parameters of the system alter the relationship between changes in reserves (or the monetary base) and changes in the quantity of money and interest rates.[1] This implies that the multipliers relating the instruments of monetary policy to real output, inflation, and employment are changed.

If the changes in the parameters of the system were known with certainty, financial innovation would pose no particular problem for monetary policy. Policy multipliers would change over time in a predictable fashion, and the proper setting for the instruments of policy would be apparent. Furthermore, if the probability distributions governing the stochastic shocks to the economy were also known, the results of rational expectations models would also hold.

Unfortunately, there is a great deal of uncertainty concerning the true structure of the economy, and the extent of parameter changes cannot be predicted accurately. Furthermore, the probability distributions of stochastic shocks also are not known and the distributions are not stationary. It is uncertainty about these matters that produces problems for monetary policy.

The introduction of interest-bearing money provides a timely example. Deregulation has produced transactions accounts for households and nonprofit institutions that pay a market interest rate. Business transactions accounts still pay no interest, but it is very likely that they will be deregulated soon. When this occurs, the entire transactions account component (currently about 2/3) of the money stock will earn a market rate of interest.

This financial innovation increases the quantity of money demanded by the public and, given the quantity of reserves in the system, shifts the *LM* curve to the left. The extent of the shift is unknowable at this time, however. Furthermore, it can be shown that if the interest elasticity of money demand does not increase substantially, this innovation will make the *LM* curve steeper. The extent of the slope change is also unknowable.

The shift and slope change in the *LM* curve require a change in the instruments of monetary policy if policy objectives are to be met. If the extent of the shift in the *LM* curve and the increase in its slope were known with certainty, monetary policy could adjust easily to the new environment. The extent of these changes are not known, however. Consider the problem in predicting the new slope of the *LM* curve. The extent to which the slope increases depends, in part, upon how depository institutions price their transactions accounts relative to the interest rates on other assets. There will be considerable trial and error in setting the pricing rule, and short-run competitive struggles have already developed that have produced a varying pricing rule. These adjustments imply that the slope of the *LM* curve will be uncertain and variable.

A possible change in the interest elasticity of money demand provides an additional source of uncertainty. When transactions accounts pay a market rate of interest, they become an active portfolio item for the public quite apart from their function as a medium of exchange. This suggests that the portion of money that is held for portfolio purposes may be highly sensitive to changes in the yield on money relative to other assets. If this occurs, and the experience with money market mutual funds suggests it will, the interest elasticity of money demand increases. A rise in the interest elasticity of money demand lessens the increase in the slope of the *LM* curve produced by deregulating the interest rate on transactions accounts. In principle, a sufficiently large increase in the interest elasticity of money demand can produce an *LM* curve that is actually flatter than the one that existed before interest rate deregulation. Uncertainty

[1] Innovation has also raised serious problems in the definition and measurement of money.

about the extent of the change in the interest elasticity of money demand is an additional reason to be unsure about the slope of the *LM* curve. It is no easy matter to conduct monetary policy in such an uncertain environment, and appeals to rational expectations arguments are of no help. There is little reason to believe that using growth in the quantity of money as the intermediate target of policy will be productive.

II

The waves of financial innovation that occurred during the last decade and a half have complicated monetary policy because they have produced unpredictable changes in the parameters of the system. The innovations themselves are primarily the consequence of government-imposed interest rate ceilings and reserve requirements. Surges of inflation, coupled with episodes of restrictive monetary policy, produced high and variable interest rates. The behavior of interest rates produced innovations that allowed market participants to circumvent the government restrictions. Attempts by the Federal Reserve to combat inflation raised interest rates and helped produce innovations. Thus, policy affected the rate of financial innovation, which in turn, affected policy multipliers. (See Donald Hester, 1981.)

The financial innovations that occurred over the last decade and a half are well known. In the 1960's, banks developed negotiable CDs, holding company debt, and Eurodollar operations as methods of circumventing Regulation *Q* interest rate ceilings. These innovations allowed banks to effectively manage the size of their liabilities. The high and variable interest rates of the 1970's spawned a number of financial innovations. Most important of these for our purposes were repurchase agreements, money market mutual funds, and interest-bearing transactions accounts.

Economists and policymakers began to be concerned about the effects of financial innovation in the mid-1970's when conventional money demand functions began to seriously overpredict the quantity of money. (See Stephen Goldfeld, 1976.) The high nominal interest rates of the time produced extensive use of repurchase agreements by large depositors, and money market mutual funds and other money substitutes began to be used by smaller depositors. High interest rates induced innovation which allowed the public to economize on its money balances to a greater extent than predicted by conventional money demand functions. Furthermore, when interest rates later declined, the technology of improved money management remained in place, and the public continued to hold less money than predicted by conventional equations. Financial innovation apparently produced a permanent downward shift in the money demand function.

The shift in money demand made monetary policy formation and evaluation difficult. For example, during the period 1973 through 1975, monetary policy appeared to be highly restrictive when one looks at the behavior of the real quantity of money or nominal interest rates. From the second quarter of 1973 through the fourth quarter of 1975, real *M*1 declined by nearly 9 percent. Short-term nominal interest rates soared and reached double digits in mid-1974. When one looks at real short-term interest rates, however, monetary policy appears to have been expansionary. In 1974 and 1975, the real federal funds rate was often negative and even when it was positive, it never reached 1.5 percent on a quarterly average basis.

The large downward shift in real money demand compensated for the decline in real money balances. Put another way, financial innovation produced an unexpected increase in velocity that prevented monetary policy from being as restrictive as it otherwise would have been. The downward shift in money demand was a surprise to economists and policymakers, and it seriously complicated monetary policy.

As a further complication, when money demand functions are reestimated using data from the 1970's, there is evidence that the interest elasticity of money demand increased relative to earlier estimates. (See Flint Brayton et al., 1983.) Financial innovation lowered transactions costs and heightened the substitutability between money and interest-bearing assets. Financial innovation produced shifts in both the location and slope of the money demand schedule. The extent and permanence of these changes were not well understood at the time. This made growth in the quantity of money a poor target for monetary policy.

The 1973–75 episode is important because it provides an example of how financial innovation can complicate monetary policy. When we turn to 1979–82, the next period of monetary restraint, the story is different, however. During the episode of the great monetarist experiment, there is no evidence of a further downward shift in money demand.[2] As with 1973–75, there was a substantial decline in real money balances, on average, during 1979–82. Nominal interest rates also rose to extremely high levels. This time, however, real short-term interest rates were very high. The peak of the real federal funds rate was higher than the peak in the *nominal* federal funds rate of 1973–75. This time the decline in real money balances was not accompanied by a downward shift in money demand.

Despite the high and variable interest rates of 1979–82, there is no evidence that financial innovation produced large unpredictable changes in the demand for money. Even though the quantity of money and interest rates experienced wide short-term fluctuations, money demand functions continued to track fairly well. With the exception of the second quarter of 1980 when credit controls helped produce a collapse in the quantity of money, prediction errors from various money demand equations were not unusually large.

A recent paper by the Federal Reserve Board's staff studies alternative money demand specifications and shows prediction errors for 1979–82 from dynamic simulations (see Brayton et al.). The paper demonstrates that specifications allowing for the growth of interest-bearing transactions accounts, so-called "other checking accounts" (OCAs), and for interest rate elasticities of money demand that vary with the level of interest rates perform better than more conventional specifications. There is no evidence, however, that any reasonable specification performed particularly poorly during the great experiment. Most specifications tend to overpredict the quantity of money in dynamic simulations but the errors are small compared to 1973–75. There is little in these simulations indicating that financial innovation was a substantial problem.

While *ex post* simulations fail to demonstrate any substantial shifts in money demand, or in interest rate and income elasticities, financial innovation did pose problems for monetary policy during the 1979–82 period. Monetary policy must be made *ex ante* and there was considerable uncertainty about how the growth of NOW and ATS accounts would affect the quantity of money demanded. Federal Reserve officials voiced concerns about OCAs in 1979 and 1980 before nationwide NOWs were authorized, and concern heightened in 1981 when NOWs became available nationwide. These concerns were shown in separate targets for *M*1-*A*, which did not include OCAs, and *M*1-*B*, which included these accounts. There was considerable uncertainty concerning the extent to which OCAs would grow as a result of shifts from demand deposit accounts versus the extent to which they would grow from shifts out of savings accounts. For a while, the Fed targeted on shift-adjusted money numbers to allow for estimates of the substitution out of savings accounts.

The concerns about OCAs were understandable. These accounts grew from 4 percent of *M*1 in December 1979 to 21 percent of *M*1 by December 1982. Choice of the appropriate monetary target depended crucially on assumptions concerning the growth of OCAs, and upon assumptions concerning the extent to which they came out of demand deposit accounts vs. savings accounts.

It is remarkable that the growth of interest-bearing transactions accounts did not produce larger errors in predicting the quantity of money. It is possible that the decline in the growth of money demand produced by continuing financial innovation that allowed the public to further economize on money balances was offset by the substitution from savings accounts into OCAs.

[2] There are insufficient data to detect any change in slope.

If this interpretation is current, the effects of money-economizing financial innovations were masked by the growth of OCAs. While NOW and ATS accounts were the only major new innovation during the period, there was rapid growth in older innovations such as repurchase agreements, including retail RPs, and in money market mutual funds. Given the unprecedented heights to which interest rates rose, it is likely that money demand functions would have seriously overpredicted money growth if OCAs had not grown so rapidly. Thus, the effects of innovations that allow the public to economize on money balances were offset by the innovations of OCAs. These offsetting effects allowed money demand equations to track fairly well. Targetting *M*1 during 1979–82 did not produce the problems encountered during 1973–75. Furthermore, the relative stability of money demand functions during 1979–82 indicates that unpredictable shifts in money demand were not the cause of the large fluctuations in money growth that occurred.

Model simulations indicate that *given* interest rates and real income, conventional money demand functions performed fairly well during the great monetarist experiment. This does not mean, however, that the Fed necessarily believed the predictions from its own models. These models predicted the increases in interest rates and the swings in interest rates that occurred. The Federal Reserve may have implicitly assumed that a large downward shift in money demand would occur. After all, model predictions of skyrocketing interest rates in 1973–75 were wrong—why believe the models this time? A distrust of models in the case of 1979–82 translates into a belief that estimated equations would shift down. The failure of the equations to shift may explain why Fed officials seemed surprised by interest rate movements, and why they apparently neglected to consider the consequences for the real economy of the high real interest rates that prevailed during the pɜriod.

III

This paper has argued that financial innovation did not produce large, unexpected movements in the quantity of money demanded during 1979–82. What does account for the wide fluctuations in money growth that occurred? Credit controls provide a partial answer for 1980, but the general answer lies with flawed operating procedures. When the Fed switched from stabilizing the federal funds rate to pursuing a target path for nonborrowed reserves, it apparently failed to appreciate the complications produced by lagged reserve accounting. Under lagged reserve accounting, the only practical reserve target is free reserves. This, in turn, made predictions of borrowing from the discount window crucially important. These predictions were not accurate and the Fed ended up assuming that borrowing was determined by a random walk. (See Paul Meek, 1982.) It can be shown that this assumed behavior of borrowing leads to open market operations that are intended to stabilize money growth, but end up destabilizing it. This destabilizing behavior of monetary policy was probably the major cause of erratic money growth during the period; not financial innovation, and not unexpected shifts in money demand.

REFERENCES

Brayton, Flint, Farr, Terry and Porter, Richard, "Alternative Money Demand Specifications and Recent Growth in *M*1," processed, Board of Governors of the Federal Reserve System, May 1983.

Hester, Donald D., "Innovations and Monetary Control," *Brookings Papers on Economic Activity*, 1:1981, 141–89.

Goldfeld, Stephen M., "The Case of the Missing Money," *Brookings Papers on Economic Activity*, 3:1976, 683–730.

Meek, Paul, *U.S. Monetary Policy and Financial Markets*, New York: Federal Reserve Bank of New York, 1982.

Lessons from the 1979–82 Monetary Policy Experiment

By MILTON FRIEDMAN*

A "monetary policy experiment" *was* conducted from October 1979 to the summer of 1982, but many commentators have misinterpreted its character and have drawn the wrong conclusions from the evidence it generated. They have termed it a "monetarist policy" and have interpreted the results as a failure of monetarism.

Though the Federal Reserve System's rhetoric was "monetarist," the actual policy that it followed was antimonetarist (see my 1983 article). And the evidence generated by the experiment strongly supports rather than contradicts the propositions that recommend a monetarist policy.

In its October 6, 1979 announcement that initiated the experiment, the Fed described its changed procedures as designed "to support the objective of containing growth in the monetary aggregates." That is indeed a monetarist objective. However, a monetarist policy involves not only targeting monetary aggregates, but also—as a major and central element—achieving a steady and predictable rate of growth in whatever monetary aggregate is targeted. By this essential criterion, the experiment was antimonetarist: as Table 1 shows, the volatility of monetary growth during the experiment was about three times as high as earlier. Indeed, monetary volatility was higher during the three years of the experiment than in any earlier three-year period since at least the end of World War II.[1] The purely rhetorical character of monetary targeting is clear also from the Fed's failure to hit its targets. *M*1 was outside the target range in six out of ten quarters, and this despite the generous width of the range between the lower and upper growth rate targets plus an annual shifting of the base to which the target growth rates were applied.

TABLE 1—STANDARD DEVIATION OF QUARTER-TO-QUARTER RATES OF MONETARY GROWTH

Ten Quarters:	*M*1	Adjusted Monetary Base
Prior to October 1979	1.59	.94
After October 1979	5.64	2.71

Many monetarists believe that slowing monetary growth will reduce inflation more promptly and at a smaller cost in terms of reduced output and higher inflation if it is announced in advance than if the slowing is not anticipated. However, their belief depends on the pre-announced slowing of monetary growth being widely believed by the relevant economic agents. Such belief was not widely present, even in October 1979, when the new policy was first announced. And the wide gyrations in monetary growth rates in subsequent months rapidly disillusioned any naive agents who initially accepted the Fed's rhetoric as a guarantee of steady and predictable monetary growth. Since there was no credible pre-announced slowing, it could not have had any of the effects attributed to such a slowing by some monetarists. Similarly, the failure of the policy to achieve such effects cannot be regarded as contradicting monetarist predictions.

*Senior Research Fellow, Hoover Institution, Stanford, CA 94305.

[1] The failure of the Fed to follow a monetarist policy is not surprising. If at the outset of the experiment, each member of the Board of Governors had been asked, "Are you now, or have you ever been, a monetarist?," not a single one would have answered "yes." In a column commenting on the October 6, 1979 announcement of the change in policy, I wrote that "those of us who have long favored such a change have repeatedly licked our wounds when we mistakenly interpreted earlier Fed statements as portending a change in operating procedures. I hope this time will be different—but remain skeptical until performance matches pronouncements" (1979, p. 39).

The real experiment was in the operating procedures adopted—the use of nonborrowed reserves as the instrument. In effect, that meant a reversion to the "free reserves" approach of the 1920's. Combined with lagged reserve requirements, the new approach produced enhanced volatility in monetary growth, and, as a consequence, in both interest rates and economic activity. As a citizen, I deplore the results, which, as I have argued elsewhere (1983), included imposing a much higher cost for the achieved reduction of inflation than was necessary. As a scientist, I am delighted to have the unintentional experimental evidence on the effect of sharp swings over short periods in the rate of monetary growth.

These sharp swings provide some evidence on two monetarist propositions: first, that different monetary aggregates move together; second, that movements in monetary aggregates produce corresponding movements in nominal income.

I. Different Monetary Aggregates

In judging the evidence on claims by monetarists that different monetary aggregates move together, it is important not to be confused by labels. The aggregates as currently defined do not correspond to the aggregates with respect to which those claims were made. They were made primarily with respect to $M1$ and $M2$ as then defined. The current aggregate labelled $M2$ is a much broader aggregate than the earlier $M2$. Indeed it is nearly identical, both conceptually and numerically, with the aggregate that Anna Schwartz and I labelled $M4$ in our *Monetary Statistics*.[2] The current $M1$ is conceptually, though in the earlier years not numerically, closer to the aggregate we labelled $M2$ rather than to our $M1$ because, like our $M2$, it includes deposits bearing interest.[3] The current aggregate that is conceptually closest to the earlier $M1$ is the monetary base.

Few if any monetarists ever recommended the use of such broad aggregates as the current $M2$ or $M3$ as monetary targets—certainly, this one did not. The closest current approximations to the aggregates they recommended are therefore the current $M1$ and the monetary base.[4]

For the five years, 1978 to 1982, the correlation between the fourth-quarter to fourth-quarter rates of growth of $M1$ and the base adjusted for reserve requirement changes is .89; of $M1$ and the unadjusted base, .63. On the other hand, the correlation of $M1$ and $M2$ is .34; of $M1$ and $M3$, .17; of $M2$ and $M3$, .17.

II. Money and Income

The asserted relation between movements in a monetary aggregate and nominal income that is relevant to current policy is about cyclical effects. In judging this proposition, a year is too long a time unit to use, especially for the period from 1980 to 1983, since it has been characterized by a succession of abnormally short cyclical phases: a six-month contraction in 1980, followed by a twelve-month expansion and then a sixteen-month contraction, interrupted by a one-quarter revival. Monetarists attribute this result to the corresponding abnormally short and exceptionally volatile gyrations in monetary growth. Moreover, monetarists have typically concluded that the lag of the rate of change of nominal income behind the rate of change of $M1$ is

[2] For example, for January 1960, our estimate of $M4$ is \$297.4 billion; the Federal Reserve Board's estimate of the current $M2$, \$298.2 billion; our estimate of our $M2$, \$208.9 billion (see our study, 1970, p. 47). For January 1983, I estimate the counterpart to the earlier $M2$ to have been \$1012.3 billion, whereas the Fed's estimate of the current $M2$ is \$2176.1 billion, or more than twice as large.

[3] In the earlier years, before the introduction of checkable deposits that paid interest, the current $M1$ was numerically the same as our $M1$, which included only non-interest-bearing assets. For example, for January 1960, the Federal Reserve Board's estimate of the current $M1$ is \$141 billion; our estimate of $M1$ is \$141.7 billion (Friedman-Schwartz, p. 47). However, for January 1983, I estimate the closest counterpart to the earlier $M1$ to have been \$373.7 billion, whereas the Fed's estimate of the current $M1$ is \$482.1 billion.

[4] Or a reconstruction of the equivalent of the earlier $M2$, something that I have not attempted.

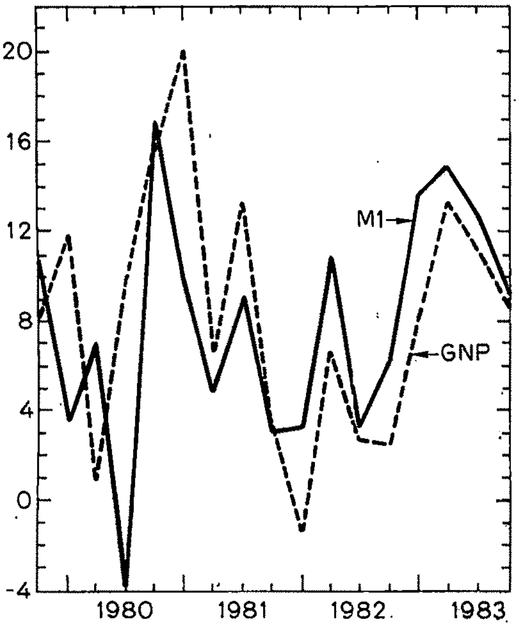


FIGURE 1. QUARTER-TO-QUARTER RATE OF CHANGE OF *M*1 AND *GNP* ONE QUARTER LATER

Note: Periods for: *M*1, 79:3–83:3; *GNP*, 79:4–83:4

TABLE 2—SWINGS IN *M*1 AND NOMINAL *GNP* ONE QUARTER LATER, QUARTERLY DATA, 1979:1 TO 1983:4

		Annual Rate of Change			
Period for *M*1	Number of Quarters	Monetary Base	*M*1	*GNP* One Quarter Later	Period for *GNP*
78:4 to 79:4	4	8.2	7.4	10.2	79:1 to 80:1
79:4 to 80:2	2	6.4	1.5	5.2	80:1 to 80:3
80:2 to 81:2	4	7.8	10.1	13.9	80:3 to 81:3
81:2 to 81:4	2	3.3	3.2	1.1	81:3 to 82:1
81:4 to 82:1	1	9.6	11.0	6.6	82:1 to 82:2
82:1 to 82:3	2	7.3	4.7	2.6	82:2 to 82:4
82:3 to 83:3	4	9.3	12.6	10.3	82:4 to 83:4

about six months on the average, in general ranging from three to nine months.

Figure 1 plots the quarter-to-quarter rates of change of *M*1 and of *GNP* one quarter later for the period of the "monetary policy experiment." From 1981 on, the relation is extraordinarily close—indeed, instead of being less close than during the earlier years, it is considerably closer. For 1980, there are significant discrepancies, almost surely attributable to President Carter's imposition and subsequent removal of credit controls. The correlation between the two series is .46 for the period as a whole; .71, eliminating the quarters affected by the credit controls.

Two things are notable about the relation between money and income in these years: first, the lag is both shorter on the average and less variable than in earlier years, second, the relation is unusually close. I believe that both are a consequence of the exceptionally large fluctuations in *M*1 growth. The effect was to enhance the importance of the monetary changes relative to the numerous other factors affecting nominal income and thereby to speed up and render more consistent the reaction.[5]

The close relation between money and nominal income is brought out in a different way in Table 2, which distinguishes the successive periods of rapid and slow growth in *M*1. The one-to-one relation between ups and downs in *M*1 growth and in *GNP* growth one quarter later is striking. There is a similar one-to-one relation between ups and downs in *M*1 and in the monetary base.

III. Money and Inflation

The long-period evidence suggests that inflation has much inertia and that the lag between money and inflation is of the order of two years. Table 3 shows that this relation has held in recent years as well. There is a one-to-one relation between movements in monetary growth, and in the *GNP* deflator two years later over successive two-year periods since 1971. The decline in inflation from 1979–81 to 1981–83 was decidedly larger than might have been expected from the decline in monetary growth. I attribute that not to a pre-announced slowing of monetary growth, but rather to the exceptional volatility of monetary growth, which increased the degree of perceived uncertainty and thereby increased the demand for money.

[5] For longer periods, annual data do provide relevant evidence on the relation between *M*1 and nominal income. For the 111 years from 1871 to 1981, the contemporaneous correlation between rates of change of *M*1 and nominal income is .72; for the 24 years from 1960 to 1983, it is .75, and rises to .86 if the *GNP* data are for a year ending one quarter later than the *M*1 data.

TABLE 3—RATES OF CHANGE IN MONEY AND IN INFLATION EIGHT QUARTERS LATER

	Annual Rate of Change Over Eight Quarters		
Period for Money	*M*1	Deflator Eight Quarters Later	Period for Deflator
71:3 to 73:3	6.9	7.4	73:3 to 75:3
73:3 to 75:3	5.1	5.5	75:3 to 77:3
75:3 to 77:3	6.4	8.2	77:3 to 79:3
77:3 to 79:3	8.5	9.2	79:3 to 81:3
79:3 to 81:3	6.2	4.8	81:3 to 83:3
81:3 to 83:3	9.2	(?)	

The increased rate of monetary growth in the 1981–83 biennium suggests that we have passed the trough in inflation and that inflation will be decidedly higher from 1983 to 1985 than it was from 1981 to 1983.

IV. Money and Real Economic Growth

The inertia in inflation and the lengthy lag between monetary change and inflation mean, of course, that the short-run influence of money on nominal income will be reflected primarily in movements in real income. And that is the case for the period under consideration. The correlation between quarter-to-quarter rates of change of *M*1 and real *GNP* one quarter later is .54 for the whole period covered by Figure 1, and .86 excluding the quarters affected by credit controls. These correlations are even higher than those for nominal income, though that has generally not been the case in the past. Again, I attribute this result to the larger amplitude and shorter duration of the recent swings in monetary growth.

V. Conclusion

The evidence generated by the misinterpreted monetary policy experiment of 1979 to 1982 is entirely consistent with the empirical conclusions about the relation between money, income, and prices that monetarists have drawn from earlier evidence. If anything, the recent evidence strengthens the case for a policy of steady predictable monetary growth. Of course, evidence for three years and a single country cannot add much to evidence provided by studies covering more than a century of experience and many countries. But the particular three years do contribute disproportionately precisely because the policy actually followed deviated so widely from the policy recommended by monetarists.

REFERENCES

Friedman, Milton, "Monetarism in Rhetoric and in Practice," *Bank of Japan Monetary and Economic Studies*, October 1983, *1*, 1–14.

______, "Has the Fed Changed Course?," *Newsweek*, October 22, 1979, p. 39.

______ **and Schwartz, Anna J.,** *Monetary Statistics of the United States*, New York: Columbia University Press for the National Bureau of Economic Research, 1970.

REFLECTIONS ON THE CURRENT STATE OF MACROECONOMIC THEORY

Reflections on Macroeconomics

By GEORGE L. PERRY*

The turmoil that has characterized macroeconomics for at least a decade originated with the inflation that emerged in the late 1960's and persisted stubbornly throughout the 1970's. The inability of the neoclassical synthesis to model inflation convincingly spawned the new classical models. Because they infer macroeconomic results more directly from principles of maximizing behavior, they appeal to some as more rigorous. Many others reject them as irrelevant because neither the microeconomic behavior that these models postulate nor their macroeconomic implications are realistic. The central postulates are that all agents are price takers and that all markets clear in the sense of Walrasian auction markets. In a dynamic, or multiperiod, context this means markets clear in rationally expected future prices. The most controversial implications are policy ineffectiveness and, in some versions, the proposition that inflation could be eliminated with little cost in real output.

As one of those who was unimpressed with these more provocative postulates and implications, I was distressed that anyone took them seriously, and that the profession became so divided over important policy issues. But it is also true that some more interesting ideas, that were originally linked with the auction-market model by Robert Lucas and other authors, might not have been developed without the controversy. One is that the rational expectations methodology should be applied in a thorough way to macroeconomic relations. Another is that private agents' behavior will depend on the rules governing the conduct of policy, or the policy regime.

I am optimistic that we are on the verge of generating more light and less heat than we have been recently. The interesting new ideas and methodology that have developed in the course of the past decade's debates will continue to be explored. But inevitably, I believe, these and other ideas will be examined within the framework of an economy that in crucial ways does not operate with auction-like markets. Both the lack of empirical success with the classical postulates and the intellectual challenge of developing a more general micro foundation to supplant the auction model are pushing in this direction. With this development, the gap between rigor and reality should narrow as researchers differ less about the basic postulates underlying macro models.

I. Auction Markets or Reality

The ideas that come from market clearing in expected future prices are second nature to economists. We are born with a talent for appreciating efficient market theory as it applies to the prices of storable commodities, financial assets, and the like. Without that appreciation, we might have grown up to be accountants or mathematicians instead. But once born, economists are raised observing that labor markets, and most product markets, do not seem to behave in this way.

The tension between observed facts and familiar theory is nowhere more evident than in attempting to integrate Walrasian market clearing and macroeconomics. It is bad science to build models that are inconsistent with the facts just because they fit a particular theory. In such a situation, it is the particular theory that should be replaced. For

*Senior Fellow, The Brookings Institution, Washington, D.C. 20036. In preparing this paper, I benefited from discussions with William Brainard, Martin Neil Baily, and Charles Schultze.

this reason, developing a better microeconomic theory to serve as an underpinning for macroeconomics has been near the top of the research agenda.

Many economists would reason that macroeconomics can go on with its business even before all the micro underpinnings are in place. It can and it has. And I agree that it should. One main stylized fact about macro behavior is that price and wage inflation have inertia, and that prices and wages are "sticky" relative to the ongoing inflation path. This fact has been employed in macro models before it had any rigorous behavioral underpinnings, and it has had greater predictive value than the alternative assumption that prices and wages are highly responsive and move promptly to clear markets, or would do so except for faulty perceptions. Macro models should continue to use some sort of sticky price assumption. But it would clearly be better to have a good behavioral understanding of it and of other macro relations.

For one thing, the sticky price assumption helps explain output and employment, but not prices themselves. Prices and wages, although sticky, are not rigid. Inflations occur and are reduced at great cost. Macroeconomics has never had a good theory of inflation, and micro underpinnings might help provide one. Furthermore, if we take seriously the idea that agents' reactions may depend on their environment, a good set of micro underpinnings could inform our thinking about how to bring about desirable changes in agents' behavior. It might provide some basis for answering whether and how the reactions of agents might change in a different stabilization policy regime. It might also provide some basis for designing and evaluating policies that are aimed more directly at changing the reactions of agents. Finally, the desire for a rigorous development of the ideas and restrictions embodied in macroeconomic relations is legitimate in itself. What is counterproductive is not that research objective, but the unsupported proposition that the Walrasian auction market offers the right postulate on which to build a rigorous macro theory. Even the notion that agents always maximize profits and utility is not sacred. Microeconomics needs to contemplate behavior that is motivated by a wider range of objectives.

Stanley Fischer (1977) and Edmund Phelps and John Taylor (1977) have tried to explain the basic fact of inflation inertia while minimizing the departure from the auction model by allowing for overlapping long-term union contracts while assuming auction-market behavior when those contracts expire. However less than 10 percent of the work force is covered by long-term contracts. The main burden of explaining average wages in these models therefore falls on how they explain the other 90 percent. The assumption that either relative wages or real wages are important in individual wage setting will help explain inertia in such models. But neither assumption fits easily with the auction-market postulates, nor do they constitute an alternative microeconomic model. Furthermore, the observed inertia of inflation predates the institution of long-term contracts. So the fundamental answer must lie elsewhere.

The burgeoning literature in microfoundations, which started formally with the implicit contract theories of Martin Baily (1974) and Costas Azariadis (1975) and has been extended by Robert Hall (1980) and many other researchers, represents a richer attempt to explain why some markets respond to changes in demand mainly with quantity rather than price variability. This literature has stressed the labor market, relying on highly damped wage responses to account for damped responses of prices. A general microeconomics should include an explicit model to explain the behavior of price-setting firms as well. Arthur Okun (1981) has provided the most comprehensive model of this type, extending contract theory in labor markets and integrating it with his model of product pricing in which "customer markets" characterize important parts of the economy. Together, Okun's labor and product markets make up what he called the price-tag economy. I will not attempt to summarize the full and realistic discussion of optimizing behavior that underlies Okun's model. It stresses that agents are concerned with intertemporal optimizing, and that the arrangements between firms and workers and firms and customers must be understood in the context of long-term relations in a world where information is costly to obtain so that shopping entails important transaction costs. As an empirical matter, the fact that prices in most markets fluctuate cyclically no more than wages suggests the customer market model in product markets is as pervasive as implicit contracts in labor markets. The customer market model has not yet been formalized and developed by others in the profession as intensely as implicit contract theories for the labor market have been. Once carefully scrutinized, I believe Okun's price-tag economy, with its equal stress on the importance of arrangements in labor and product markets, will be the heart of the general microeconomics that we are seeking.

One important point that emerges from the research by Okun and others on micro foundations is that although firms, workers, and consumers operate under implicit contracts and customer market arrangements, their doing so does not imply macroeconomic efficiency. Even along a steady-growth equilibrium, price generally exceeds marginal cost in customers markets. And when a shock displaces the economy from such a path, the macroeconomic consequences are inefficient even if firms are making their most advantageous short-run response in their own labor and product markets.

Basically, optimal arrangements at the microeconomic level are the best ones that firms and workers can make given the macroeconomic environment in which they expect to operate. But they would be better off if macroeconomic disturbances did not occur, or if they could be attenuated by policymakers. Firms and workers do their best to prepare for waves. But they would prefer smoother seas.

There is a lingering suspicion in some quarters that, despite appearances, experienced outcomes must be Pareto optimal, because if they were not, different arrangements would have been made among agents to freely capture the wasted utility. Such a view precludes any meaningful attempt to build a more useful microeconomic underpinning for macro models because, by assumption, any implicit contracts or other market arrangements developed by agents would serve merely as a veil behind which the underlying conditions of Pareto optimality continued to operate. There are rejoinders to that view based on the inability of private agents to coordinate their market-clearing responses (Robert Gordon, 1981). Individual agents cannot profitably move prices to ensure full employment in the way that all private agents could if they acted in concert. This line of reasoning provides one way to view the role of policy: it generates the macroeconomic income effects that would be forthcoming if private agents could coordinate to vary prices.

Another answer to the Pareto optimality objection lies in broadening our notion of what motivates private agents. In this sense, the objection misses the main point of the price-tag economy. Being able to operate in a world of sticky nominal wages and prices, with the arrangements among agents that lead to them, confers intertemporal benefits to agents. The benefits of operating with these arrangements outweigh the narrowly defined utility gains that might be available if the arrangements were abandoned in favor of alternatives with volatile wages and prices. And agents cannot have it both ways. It is a contradiction to model firms as reaping the benefits of operating in the price-tag economy, and then allowing them to vary prices so as to clear markets in response to demand shocks.

The microeconomics of a price-tag economy does not deny that departures from full employment may set in motion private actions that tend to restore it. A general microeconomics that explains employment variability does not predict frozen wages and prices. It should explain the moderate degree of inflation variability that is observed. And it should allow for the possibility that the implicit contracts and other arrangements that define behavior in a price-tag world in normal times can themselves be altered under sufficiently extreme conditions, or possibly in response to important changes in the policy regime, or in other dimensions of the economic environment.

We have recently observed historic changes from previously existing relations between many firms and their workers. In the airlines, arrangements are changing from union to nonunion. In major durable goods industries, wage concessions, which were only a curiosity in the past, became an epidemic in the past two years. In isolated cases, workers are for the first time accepting some profit-sharing and equity positions in exchange for wage concessions. It is significant that in all these instances, major changes have come in response to desperately hard times by individual firms. They have not been distributed widely across the economy as would be expected if they were a response to expectations about general economic conditions, or to a change in the policy regime.

II. Rational Expectations in Macro Models

The notion that expectations should be modelled as rational is a compelling one in comparison with *ad hoc* alternatives. But it does not settle what it is that agents have meaningful expectations about, how their behavior is affected by them, and how policy changes relate to the expectations agents care about. I believe that exploring these issues in order to incorporate the rational expectations methodology into macro models in an appropriate way is another direction in which macro modelling can progress. In important ways, these issues will depend on whether the auction-market postulates are assumed, or whether a more general micro model, characterized above by the price-tag economy, is used.

One issue concerns what decisions depend on expectations. Decisions are inherently intertemporal and any economic agent would like to know the future. But the importance of such information, its reliability, and the cost of acquiring it will all differ among agents. For example, if forecasts of changes in wage inflation are poor and if wages will be adjusted again next year, the value of using a forecast may be more than offset by what doing so would add to the contentiousness of wage setting. In this case, rational agents may deliberately use the current rate of wage inflation in place of any forecast. If they do, using rational expectations in modelling wage changes will exaggerate the effect of expectations on inflation.

The real force of applying the rational expectations assumption comes when policy changes are contemplated. What Brainard calls the expectations multiplier will depend both on how expectations affect agents' behavior and on how strong an effect policy changes have on those expectations. Modelling expectations dependent on simple descriptions of policy variables, such as the money supply, may further exaggerate the expectations multiplier that will be estimated to come from a policy change. Even where expectations about economic performance are important to agents, if the link between such an instrument and economic performance is uncertain, the instrument is unlikely to form the main basis for expectations. This argues for modelling directly the things about which agents are likely to hold expectations. These micro concerns probably aggregate to nominal or real *GNP* for some decisions, and perhaps to the inflation rate for others.

The inertia of prices in a price-tag economy is also a reason to believe that expectations of those prices is not central to the output decisions of firms, although expectations about other things may be important. We commonly observe that price-setting firms make substantial efforts to forecast quantities, receive market signals about demand through changes in inventories and orders, and respond at least in part by adjusting quantities. To cite the prototypical industry that most people are aware of, automobile producers make substantial efforts to predict unit sales and adjust production levels as their sales fail to match expectations, and as their expectations of future sales change. Firms operating in this way can be seen as optimizing in Okun's customer markets.

The recent macro modelling by Peter Neary and Joseph Stiglitz (1983) and by Olivier Blanchard and Jeffrey Sachs (1982) exploits rational expectations about quantities in a way that usefully generalizes the application of the rational expectations methodology. These papers focus on intertemporal decision making by agents and the application of rational expectations constraints to their behavior. They demonstrate policy effectiveness under a range of conditions and formally address questions of recent interest such as the effect of policy anticipations. Just as John Muth's (1961) original examples and those of Lucas and others who followed him were appropriate for the auction-market postulates, this recent work suggests how rational expectations might be useful in the more general micro model.

III. Stabilization Policy

The possibility that private behavior would be altered by policy changes—the Lucas critique—is one of the lasting contributions of the past decade's debates. In the general issue of policy design, it has been most important in focusing attention on the possible importance of different policy regimes. Alternative regimes might be described as overactivist, just right, or underactivist. The one polar result that has received renewed attention is the optimality of a fixed rule without feedback—or studied neglect. Historically that rule has led to suggestions for a fully committed policy of constant growth of the money supply, unchanging fiscal policy such as might be dictated by a balanced budget amendment, or other such devices for precommitting to a fixed policy. In recent years it has received formal support as an implication of rational expectations models in an auction-market world.

I do not want to dwell on that polar case, although it has been useful in sharpening the discussion of stabilization issues. If behavior of the private sector always kept the economy on its full-employment equilibrium path, or if departures from that path were simply the result of unanticipated variations in policy, studied neglect would be optimal. The general microeconomic model does not produce that kind of macroeconomic behavior. It forces us to consider stabilization policy in a world where shocks can move the economy from such a path, and where all kinds of policy changes can have real effects.

Studied neglect may still offer certain benefits, but it will also entail clear costs and risks. In evaluating that and alternative rules for conducting policy, we do need to consider that how the economy performs may affect the functions describing how agents behave. As a general proposition, there is probably no limit to how far one could push that idea, at least distantly relating the laws and institutions of the society to its economic performance. The Great Depression brought on trade protection, legislation strengthening unions, and the welfare state with consequences for agents' behavior. Inflation in the postwar years popularized price indexing in long-term contracts and spurred deregulation of the financial sector. Although we can understand such developments, it is hard to allow for these and more subtle effects in building macro models.

Classical rational expectations models claim to make the right allowances. And for an important set of issues, they make a crucial point. When monetary policy went to restrictive money supply rules in the fall of 1979, a rational expectations model warned that bond prices, set in efficient auction markets, might depart from the predictions of term-structure equations that were blind to the advent of the new policy regime. However, the relevance of that methodology for modelling regime changes is not clear when applied to other economic issues. The particular problem that Lucas (1972, 1973) originally considered and that remains a central concern of macro theory is whether the economy's unemployment-inflation performance can be improved by a suitable choice of policy rule. Lucas' results are tied to the auction-market postulates. A general microeconomic theory does not yet exist to deal adequately with that question, and for some time we may have to look to experience and more casual model building for guidance.

It seems to be true that a policy regime dedicated to maintaining high employment gradually finds that more inflation accompanies any given level of unemployment. The 1960's are the prime example. Inflation climbed steadily under the economic developments of the last half of the decade, establishing a higher norm rate for wage increase. This higher wage norm, which for simplicity we can think of as stable, then characterized the decade of the 1970's, with cyclical developments and the two oil price shocks leading to variations in actual inflation around the high norm. Now, only after a deep and prolonged recession, the wage norm is apparently coming down again.

There is more than one plausible model of the events I have just described, and no clear verdict on how to improve performance in a lasting way. The wage norm doubtless is closely related to inflationary expectations. What it takes to change the norm is the important question. Does it require the actual experience of severe unemployment to lower it? Or is there some shortcut available through the clear announcements of policy intent? One prediction of new classical models was that credible disinflationary policy would end inflation more promptly and with a smaller loss in output than historical data, fit to a more employment-oriented regime, would predict. Empirical evidence from the great recession we have just been through shows the disinflation was not especially prompt (see my 1983 paper). On my interpretation, that evidence also shows that eventually the norm did shift down and the improvement in inflation will now prove greater than a purely cyclical Phillips curve model would predict. That, too, is a prediction of more than one model.

Charles Schultze (1981) has provided empirical evidence on inflation and output for the United States from before the start of this century, in which he separates the cyclical response of inflation from norm shifts. In this period, all kinds of changes have occurred in the economic structure of the country, including big ones such as its industrialization, waves of immigration, the rise of unions, and the growth of the welfare state; and smaller ones such as the introduction of three-year labor contracts and the rise, fall, and then rise again of price indexing in those contracts. There have been vastly different policy regimes over this period, including regimes of implicit, if not studied, neglect. What is remarkable and, to me, surprising, is that economic performance has differed mainly in the stability of real output and the frequency of recessions. There is no evidence that the behavior of private agents varied so as to produce more price flexibility under some regimes than under others, except in periods of war and in the Great Depression. At other times, the short-run division of *GNP* between increases in prices and in output has been little changed. Gordon (1982a), using a different model, finds some increase in the effect of lagged on current inflation in the postwar period, but an unchanged short-run response.

The fact that it took wars and the depression to produce different behavior is something to think about. Along with Thomas Sargent's (1980) study of hyperinflations, it suggests that maybe only abnormal conditions will produce abnormal behavior. If so, do we even want to look for a way to get more price flexibility? The substantial consistency of behavior observed under normal conditions with widely different stabilization regimes suggests that the institutions and arrangements underlying price inertia have transactional properties that are highly desirable to economic agents. If greater price flexibility can only be bought by changing those institutions and arrangements, we might not want even to attempt it.

In cross-country comparisons, Gordon (1982b) found some differences between U.S. Phillips curves and those for Great Britain and Japan. Much further work along these lines might identify the characteristics that make for cross-country differences, and whether there was anything transferable that we could learn from each other. It seems doubtful that the conduct of stabilization policy is the key. The most alarming fact one observes from a casual look abroad is that unemployment has risen for ten straight years in Europe as a whole. Conservative policies dedicated to price stability in Germany throughout most of this period and in the United Kingdom for the last half of it have not produced inflationless prosperity.

It must be true that a policy regime dedicated to maintaining price stability can succeed. But there is no evidence that it can do so without averaging less stability in real outcomes and higher levels of wasteful unemployment than the employment-oriented regime. The basic idea is that it makes a difference what surprise policy responds to. If it focuses mainly on offsetting surprising weakness in output, it will generate reactions that gradually generate more inflation. If it focuses on offsetting price level surprises, it will average less employment. This suggests that, so long as we confine ourselves to the traditional tools of stabilization policy, we may have to choose among middle-runs in which average inflation and average unemployment are inversely related. But a deeper understanding of what behavior underlies macro performance in a price-tag microeconomy might produce a happier verdict, or point to other policies for improving performance. This continues to be the central challenge for macroeconomic theory.

REFERENCES

Azariadis, Costas, "Implicit Contracts and Underemployment Equilibria," *Journal of Political Economy*, December 1975, *83*, 1183–1202.

Baily, Martin N., "Wages and Employment under Uncertain Demand," *Review of Economic Studies*, January 1974, *41*, 37–50.

Blanchard, Olivier J. and Sachs, Jeffrey, "Anticipations, Recessions and Policy: An Intertemporal Disequilibrium Model," *Annales de L'Insee*, 1982, *47–48*, 117–44.

Fischer, Stanley, "Long Term Contracts, Rational Expectations, and the Optimal Policy Rule," *Journal of Political Economy*, February 1977, *85*, 191–205.

Gordon, Robert J., "Output Fluctuations and Gradual Price Adjustment," *Journal of Economic Literature*, June 1981, *19*, 493–530.

______, (1982a) "Wages and Prices are Not Always Sticky: A Century of Evidence for the U.S., U.K. and Japan," Working Paper No. 847, National Bureau of Economic Research, January 1982.

______, (1982b) "Why U.S. Wage and Employment Behavior Differ from That in Britain and Japan," *Economic Journal*, March 1982, *92*, 13–44.

Hall, Robert E., "Employment Fluctuations and Wage Rigidity" *Brookings Papers on Economic Activity*, 1:1980, 91–123.

Lucas, Robert E. Jr., "Expectations and the Neutrality of Money," *Journal of Economic Theory*, April 1972, *4*, 103–24.

______, "Some International Evidence on Output-Inflation Tradeoffs," *American Economic Review*, June 1973, *63*, 326–34.

Muth, John F., "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–35.

Neary, J. P., and Stiglitz, J. E., "Toward a Reconstruction of Keynesian Economics: Expectations and Constrained Equilibria," *Quarterly Journal of Economics*, Suppl. 1983, *98*, 199–228.

Okun, Arthur M., *Prices and Quantities: A Macroeconomic Analysis*, Washington: The Brookings Institution, 1981.

Perry, George L., "What Have We Learned About Disinflation?," *Brookings Papers on Economic Activity*, 2:1983, 587–602.

Phelps, E. S. and Taylor, J. B., "Stabilizing Powers of Monetary Policy Under Rational Expectations," *Journal of Political Economy*, February 1977, *85*, 163–90.

Sargent, T. J., "The Ends of Four Big Inflations," Working Paper No. 158, Federal Reserve Bank of Minneapolis, 1980.

Schultze, Charles, L., "Some Macro Foundations for Micro Theory," *Brookings Papers on Economic Activity*, 2:1981, 521–76.

Autoregressions, Expectations, and Advice

By THOMAS J. SARGENT*

Macroeconomists have long spent most of their time observing and interpreting aggregative economic time-series. Recent progress has involved formalizing and standardizing this activity into one of estimating and interpreting vector autoregressions. Vector autoregressions provide a convenient way of summarizing the second moments of time-series data, and conform naturally with the recursive decision theory that is associated with stochastic dynamic rational expectations models. In interpreting vector autoregressions, fundamental differences exist between users of rational expectations econometrics and users of "atheoretical" or "uninterpreted" vector autoregressions in the style of Christopher Sims (1980, 1982) and Robert Litterman (1980, 1981).[1] This paper describes these differences, and uses dynamic rational expectations theory to describe strong points of each approach. That theory can be used as forcefully to support Sims' style of more or less uninterpreted vector autoregressive empirical work as it can be to justify the "fully interpreted" or structural vector autoregressive empirical work practiced by rational expectations econometricians.

My purpose is to allude to specific formal models that exist in the literature on rational expectations models, and that can be used to support Sims' actual econometric practices and many of his remarks. To date, Sims' arguments on the points in this paper have been informal, and have not been made in the context of concrete models. Therefore, in advancing the arguments, I have taken the risk that I might miss what Sims has in mind. In fact, in oral remarks, Sims has told me that my interpretation fails to capture his thoughts, and should be labeled as my own argument. Nevertheless, I continue to attribute to Sims a line of argument which he disowns.[2] I do so because the argument that I attribute to him was gathered during my reading of Sims' work and is not my own creation. Sims has told me that he agrees that the argument in the text has the virtue of disposing of often-encountered arguments to the effect that users of vector autoregressions in the style of Sims and Litterman must be regarded as ignoring dynamic economic theory.

I. Vector Autoregressions and Dynamic Macroeconomics

Robert Lucas' (1976) critique of econometric policy evaluation procedures concerned proper ways of interpreting and manipulating vector autoregressions. Lucas observed that it violated dynamic economic theory with purposeful agents, as standard procedures then did, to change one equation representing government policy actions in a

*Department of Economics, University of Minnesota, Minneapolis, MN 55455 and Federal Reserve Bank of Minneapolis. I thank Neil Wallace, Bennett McCallum, Robert Townsend, Robert E. Lucas, Jr., Robert Litterman, Bruce Smith, and Christopher Sims for helpful comments.

[1] The philosophy of rational expectations econometrics is described in the introductory essay of Robert Lucas and myself (1981) and by my 1981 article. The approach of Sims and Litterman is described in Sims (1980), Litterman (1980, 1981), and Thomas Doan et al. (1983).

[2] In constructing my unauthorized interpretation of Sims, I have selected and emphasized some themes in his writings, and have deemphasized and deleted others. My intention in doing so has been explicitly to use dynamic rational expectations theory to present a defense of Sims' actual econometric practices. Among Sims' words with which my interpretation might be inconsistent are such statements as those found in his 1982 article: lines 7–13, p. 108; lines 27–29, p. 123; and lines 3–13, p. 151. Whether my interpretation is consistent with Sims' words hinges on the meaning of the terms "policy analysis" and "useful." If they refer to giving advice in choosing government policy actions, my interpretation does not apply. It does apply if the terms refer to making probability statements about the consequences of alternative realizations of policy actions on the basis of the historically estimated probability structure.

vector autoregression while holding fixed the remaining equations, many of which describe private agents' decisions. Of the procedures that Lucas criticized, the most sophisticated explicitly posed an optimal control problem for the government as the way of finding the best equation for government policy variables, holding fixed the remaining equations in an estimated vector autoregression. In such a control problem, the object of choice is a rule or regime for the government, and the predicted outcome of that choice is a new and improved probability structure for the economy. Lucas observed that dynamic economic theory implies that in general all of the equations in the vector autoregression can be expected to change with such a change in regime, not just the equations describing the government policy.

One constructive response to Lucas' observation has been an ambitious research program to build workable dynamic rational expectations models and methods of estimation that can be used to predict how all of the other equations of a vector autoregression will change when one equation describing a government policy variable is hypothetically altered.[3] The goal of this rational expectations econometric program remains ultimately to search for rules for government policy variables that are predicted to imply the most desirable vector autoregression for the economy. The intention is thereby to obtain good practical quantitative advice for formulating new strategies for government actions in the years beyond the sample period.[4]

The most telling criticism of rational expectations econometrics has come from Sims (1980, 1982) in a sequence of remarks about appropriate ways of estimating and utilizing vector autoregressions. While accepting the theoretical observation underlying Lucas' critique, Sims challenges rational expectations econometrics, and does so by appealing to the very same general body of dynamic economic theory that Lucas used. Sims' vision and rational expectations econometrics are based on different models of the economy. To begin, it is necessary to clarify what is meant by a model economy.

A model economy consists of a collection of agents arranged in a particular way over time and space; a description of agents' endowments of and preferences for goods; a technology for converting goods into one another, possibly at different points in time and space; and a mechanism for arranging agents into coalitions or institutions, and for coordinating decisions both within and across coalitions. This conception of an economy is so broad that it leaves open whether the coordination mechanism is a Walrasian one, or an alternative one that, when compared with a Walrasian mechanism, seems to constitute a "disequilibrium."[5]

With a given mechanism, the economy can be viewed as the solution of a dynamic game.

[3] Contributions that share this goal, while differing in some technical details, are represented by John Taylor (1979, 1980, 1982) and Lars Peter Hansen and myself (1980), by Finn Kydland and Edward Prescott (1982), and by Hansen and Kenneth Singleton (1982).

[4] In the brand of rational expectations econometrics that I am describing, the historical time-series are supposed to have been generated as the solution of a dynamic game whose outcome can be improved upon. This can occur in a variety of rational expectations equilibria in which suboptimal government behavior in conjunction with nonneutralities prevent rational expectations equilibria from being Pareto optimal. Some useful examples are studied by Zvi Eckstein and Martin Eichenbaum (1984a, b). In such contexts, the computationally convenient equivalence between a rational expectations equilibrium and the solution of the social planning problem, which Lucas and Prescott (1971) have fruitfully exploited, fails to occur. Such nonoptimal rational expectations equilibria must be computed by methods other than those used by Lucas and Prescott (see Kydland and Prescott, 1977; Dennis Epple et al., 1984; Paul Romer, 1982; and Charles Whiteman (1983)). Notice that Willem Buiter's (1980) characterization of some rational expectations work in macroeconomics as being "the economics of Dr. Pangloss" does not apply to the line of work that I am summarizing under the category of rational expectations econometrics. Rather, it is Sims' criticism of rational expectations econometrics that comes closer to resting on the view that "we live in the best of all possible stochastic processes."

[5] Robert Townsend (1983a) describes a model economy with purposeful price-setting agents. For a particular game that he sets out, a Walrasian equilibrium is a solution. Presumably, settings exist in which games specified in terms of the same primitive objects as Townsend's have non-Walrasian solutions.

In a dynamic game, the strategy of each agent depends on the strategies chosen by "nature" and the other agents in the system. Such strategic interdependence is the reason that when an equation in a vector autoregression describing one agent's strategy is hypothetically altered, other equations should also be expected to change. The "rational expectations revolution" in macroeconomics consists of a broad collection of research united mainly by an aim to respect the principle of strategic interdependence.

This principle is respected and heavily exploited both in rational expectations econometrics and in Sims' criticism of that program. The difference between Sims and rational expectations econometrics involves both the specification of the dynamic game that was imagined to have been played during the historical sample period, and also the dynamic game that is imagined to occur after the sample period, when the analyst is contemplating an intervention.

II. Rational Expectations and Fully Interpreted Vector Autoregressions

Rational expectations econometrics proceeds on the supposition that the dynamic game that was being played during the estimation period is different than the one to be analyzed, and whose playing is to be recommended in the future. Rational expectations econometrics typically assumes that during the historical sample period, private agents' decisions solved constrained stochastic intertemporal optimization problems. Among private agents' constraints are laws of motion for government policy variables. Usually, these government policy variables are posited to follow arbitrary, more or less general, stochastic processes. The econometrician specifies parametric forms for preferences, for technologies, and for stochastic processes for government policy variables and other exogenous variables. The econometrician next imposes the hypotheses that private agents have rational expectations and behave purposefully, and that there is a given mechanism assuring consistency among various individuals' decision strategies and their perceptions (a rational expectations competitive equilibrium is one tractable and commonly used such mechanism). It is then possible to define a mapping from the free parameters of agents' preferences and constraints to the implied theoretical probability distribution of all of the variables in the model. The second moments of this probability distribution can be completely summarized as a theoretical vector autoregression. Given this mapping, estimation can proceed using the philosophy of maximum likelihood or generalized method of moments, each of which selects free parameters of the model so that the theoretical probability distribution (or vector autoregression) matches the empirical one as closely as possible.

Once estimates of the free parameters of agents' objectives and constraints have been obtained, the aim is to use them to analyze how the economy would behave under hypothetical strategies for setting government policy variables that are different from the one evident in the sample period. A systematic search is to be conducted for the government strategy that optimizes the performance of the economy as a stochastic process in the following particular sense. An intertemporal objective function is posited for the government, in some cases the utility functional of the representative private agent in the model. A dynamic game is formulated in which the government is dominant and is imagined to choose a strategy to maximize its criterion, taking into account the reaction of optimizing private agents' strategies to the government's choice of strategy. A solution to this dynamic game will in general be a stochastic process for the economy that differs from the one in place during the sample period.[6]

The piece of this stochastic process that describes the government's policy decisions is what the rational expectations econometrician is prepared to recommend for policy; the entire vector stochastic process is his prediction about how the economy would behave were his policy recommendations to

[6] An exception to this statement occurs when the government policy rule was optimal during the sample period.

be adopted. This hypothetical stochastic process will have a vector autoregressive representation, which will generally differ from the one in the sample period.

This setup envisions that government behavior may have been guided by different principles during the sample period than are hypothesized to guide it during the future. During the sample period, there is permitted to be an asymmetry between the principle guiding the government's strategy, which is arbitrarily (if generally) specified, and that guiding private agents, which is purposeful. However, for computing the optimal government policy strategy for the future, this possible asymmetry of behavior is removed, and both private agents and the government are supposed to be purposeful. (*Some* such difference in government behavior between the sample period and the hypothetical future *must* be posited by anyone recommending a change in the government's strategy.)

III. Sims' Challenge

I interpret Sims as objecting to the assumed asymmetry between private agents' behavior and government behavior during the estimation period. Sims' view is that the asymmetry should be eliminated by assuming that government agents as well as private agents have behaved as rational expectations intertemporal optimizers during the sample period.[7]

This view has definite consequences for how the time-series are to be interpreted in terms of deep parameters of preferences and constraints during the estimation period, and would require modifications of most existing methods of rational expectations econometrics.[8] Further, if the very same dominant player dynamic game is imagined to be played during and after the estimation period, the presumption is that the same stochastic process will hold both during and after the estimation period. It is then of no interest to analyze a change in the government strategy or "regime" because government behavior is posited to be determined by the same purposes and constraints after the estimation period as before it.[9]

On the one hand, this view implies that, while it could be extended to include parameters for government agents, the rational expectations econometrics pursuit of the deep parameters of agents' preferences and constraints is in itself of little value.[10] Interventions in the form of changes in government strategies are supposed not to occur; but the ability to analyze such interventions is a major motive of rational expectations econometrics. On the other hand, this view implies that for forecasting, it is useful to estimate vector autoregressions that are left uninterpreted in terms of parameters of agents' preferences and constraints. Such vector autoregressions can be used to produce linear least squares forecasts of the future vector of variables given past values. They can also be used to predict the future of some variables, conditioned on particular assumed paths over part of the future for some government policy variables.[11]

[7] "...it is not clear that the existing pattern of policy in most countries, in which there is weight given to stabilization of inflation, employment, and income distribution, is very far from an optimal policy" (Sims, 1980, p. 14).

[8] Many of these modifications are described in Epple et al. Note that a rational expectations econometrician imposing one of their games *E* or *F* during an estimation period would automatically be led to a position close to Sims' when it comes to giving advice for an improved regime outside the sample.

[9] "Accurate projections can be made from reduced form models fit to history because it is not proposed to change the policy rule, only to implement effectively the existing rule" Sims (1980, p. 13).

[10] If the overidentifying restrictions imposed on a vector autoregression by a rational expectations dynamic game are approximately true, imposing them could be defended on the same instrumental grounds that Doan et al. and Litterman (1980) give for imposing restrictions not based on explicit economic theories.

[11] Sims states (1982): "...effects of policy actions that affect expectation-formation mechanisms can be correctly evaluated with models that are reduced form in the sense that they do not explicitly display parameters of the expectation formation mechanism" (pp. 115–16); "They [the procedures] take account of policy endogeneity by generating true conditional projections, given specified paths for policy variables" (p. 150); "...a valid reduced form will make relatively precise conditional projections for the effects of policy actions or sequences of actions that are close in form to what has been observed historically" (p. 118).

Both of these kinds of forecasts are to be made in a way that respects the assumption that the same stochastic process governs the vector of variables during and after the estimation period. In particular, both kinds of forecasting exercises are formulated mathematically as projections of unknown variables on known variables, using the estimation-period vector autoregression as the probability model. These are the only kinds of forecasting exercises that are of practical interest, given Sims' conception of the game. These forecasting exercises respect the principle of strategic interdependence underlying Lucas's critique, because they never involve hypothesizing an altered strategy for the government outside of the sample period.

Closely related to his challenge of the estimation period-policy prescription period asymmetry in rational expectations econometrics, Sims has disputed the appropriateness of the commonly formulated dominant player dynamic game as one in which the government is imagined to search for a regime or strategy to be used into the indefinite future. In such a formulation, the government is an infinitely lived agent that is imagined to attain beneficial outcomes now by binding itself to state-contingent future actions. These beneficial effects come through "expectations effects," or, more precisely, through the workings of the principle of strategic interdependence.

Sims doubts the applicability of a setup in which an infinitely lived government is imagined to choose among infinitely lived "regimes."[12] Instead, he can be interpreted as urging that the government be thought of as consisting of an intertemporal sequence of finitely lived agents. There is a sequence of administrations, each of which cannot commit its successor even though current options are influenced by the public's speculation about what successor "governments" will do. Typically, it is posited that each administration is followed by successors with similar objective functions.[13]

In such a formulation, government agents are imagined to be dominant vis-à-vis private agents, and to take into account the effects of their current actions on future actions of private agents. However, because they cannot commit their successors, in optimizing they disregard the effects on private agents of those future government actions that are beyond the control of the current administration. This sequential setup is designed to reflect a reality in which government policy emerges from a succession of personalities within a succession of administrations, each lacking power to commit its successors.[14] In such games, there is no sense in which a policy regime is chosen once and for all. Instead, the stochastic process for government policy is determined by the purposeful behavior of a temporal hierarchy of agents, each of whom influences or controls only a few periods of actions.

IV. Contradictions

Neither of the two above approaches is free of philosophical difficulties. The rational expectations econometrics position is exposed to the following internal contradiction. Suppose that the free parameters of private agents' preferences and constraints have been estimated during an estimation period, and then are used to calculate a new and im-

[12] Sims (1982): "Yet in practice macroeconomic policymaking does not seem to be this sort of once-for-all analysis and decision. Policymakers ordinarily consider what actions to take in the next few quarters or years, reconsider their plans every few months, and repeatedly use econometric models to project the likely effects of alternative actions" (p. 109); "But permanent shifts in policy regime are by definition rare events. If they occurred often they would not be permanent" (p. 118).

[13] Sims (1982): "Policy is not made by a single maximizing policymaker, but through the political interaction of a number of institutions and individuals. The people involved in the process, the nature of the institutions, and the views and values of the public shift over time in imperfectly predictable ways" (p. 110); "...disputes about the optimal rule are no more important in principle than disputes about how to implement the existing 'rule' as it emerges from existing institutions and interests" (p. 139).

[14] Such games have been analyzed by Abderrahmane Alj and Haurie Alain (1983) and Epple et al. For a general reference on dynamic games, see Tamer Basar and Geert Olsder (1982). Epple et al.'s "game F" comes close to embodying the conception that Sims seems to have in mind.

proved strategy for government policy in the future. On the one hand, if this procedure were in fact likely to be persuasive in having the policy recommendation actually adopted soon, it would mean that the original econometric model with its arbitrarily specified rules for government policy had been misspecified. A rational expectations model during the estimation period ought to reflect the procedure by which policy is thought later to be influenced, for agents are posited to be speculating about government decisions into the indefinite future. On the other hand, if this procedure is not thought likely to be a source of persuasive policy recommendations, most of its appeal vanishes.[15]

Sims' vision is subject to the observation that, smacking of determinism as it does, it lends limited interest not only to the estimation of "structural" or "fully interpreted" vector autoregressions, but to any vector autoregression whatsoever.[16] In dynamic decision and game theory, the relevant choices are among different stochastic processes. In these terms, an analysis that maintains the assumption of a fixed probabilistic structure permits no policy advice to be given or choices to be delineated. Vector autoregressions become tools for passive, intervention-free forecasting of various types.

V. Positive and Normative Economics

The view that I have attributed to Sims subverts normative economics, and does so by embracing a normative or optimizing theory of government behavior as a positive theory of how the government has actually behaved. This method of understanding past government policy leaves no room for improving government policy in the future. In economics and other social sciences, there are many examples that exhibit the tension between having a theory that, on the one hand, can fully explain existing institutions and events in terms of purposeful behavior, and on the other hand, can be used to give advice for improving things. For example, this tension exists between the theory of the firm and operations research.[17] The recent work on vector autoregressions and dynamic rational expectations theory manifests this tension in a new and dramatic context, but the tension itself is an old one.

Sims has ample company among recent macroeconomists in using the hypothesis of optimizing government agents to explain observations on government behavior. Two examples will illustrate this. First, there is Lucas and Nancy Stokey's model (1983) of the optimal time structure of government deficits, and of the term structure of government debt. Their model gives the prediction that in response to temporary large current account government expenditures, such as those associated with wars, the government should run temporary deficits, issuing state-contingent interest-bearing securities to finance the deficit. The motive for running deficits is to minimize the distortions from taxes by smoothing their occurrence over time. The Lucas-Stokey normative results can be regarded as a positive theory of government finance that seems to work well for the advanced countries over much of the last 200 years. Under the gold standard, wars were typically accompanied by temporary suspensions of convertibility into specie of government notes, with the probability of subsequent resumption at par depending directly on the prospects for military victory. This suspension pattern governs European and U.S. data from 1789 to 1928, and can be regarded as a mechanism designed to ap-

[15] In formal work, this contradiction is evaded by regarding analyses of policy interventions as descriptions of different economies, defined on different probability spaces. The mental comparison is among economies identical with respect to private agents' preferences and technologies, but differing in government policy regime. The contradiction mentioned in the text surfaces when attempts are made to put these formal results to practical use by speaking of regime changes that are imagined to occur suddenly in real time. My article with Neil Wallace (1976) described a version of this contradiction.

[16] The issue here also concerns Gabriel Garcia Marquez (1971, 1983) and D. M. Thomas (1982), who describe settings in which individuals have access to more or less imperfect foresight, but exercise few effective choices.

[17] This example was suggested to me by Robert E. Lucas, Jr.

proximate a Lucas-Stokey optimal fiscal policy.[18]

A second example is John Bryant and Wallace's thesis (1984) that the observed denomination structure of government debt (interest-bearing securities are typically issued only in large denominations, while non-interest-bearing ones are available in small and large denominations) is to be understood as the result of the government's effort to price discriminate between different classes of portfolio holders, so as to raise seignorage and finance expenditures in an efficient way.[19,20]

VI. Concluding Remarks

I find persuasive the preceding defense of empirical work in the style of Sims and Litterman. While that work is "atheoretical" in terms of restrictions that it actually imposes on estimated vector autoregressions, it has foundations in terms of a deep and consistent application of rational expectations dynamic theory.

Nevertheless, my own response to the tensions highlighted by Sims' arguments is to continue along the path of using rational expectations theory and econometrics.[21] This response is based partly on the opinion that existing patterns of decentralization and policy rules can be improved upon. This opinion comes from the perception that unfinished and imperfect theories have been believed and acted upon in designing and operating institutions. Dynamic rational expectations models provide the main tools that we have for studying the operating characteristics of alternative schemes for decentralizing decisions across institutions and across time. I believe that economists will learn interesting and useful things by using these tools.

[18] A closely related example is Robert Barro's work (1979) using a theory of dynamic optimal taxation to generate restrictions on time-series observations on government finance.

[19] It is noteworthy that during 1922–23, when the Soviets were intent on extracting seignorage, two government currencies circulated simultaneously, the large denomination, stable valued chervonetz, and the small denomination, rapidly depreciating Soviet ruble (zovzhnak). Keynes remarked about the ingenuity of the Soviets in devising clever ways of extracting seignorage.

[20] Another example is Townsend's (1983b) theory of economic development and financial intermediation.

[21] I am unaware of an alternative approach to Sims or to rational expectations econometrics that avoids these contradictions and tensions. Even informal descriptions and analyses of historical and contemporary economic events inevitably involve somehow tentatively resolving these tensions, whether explicitly or implicitly. The nature of this tentative resolution, which cannot be totally satisfactory on philosophical grounds because of the contradictions mentioned above, often governs ones interpretations in an important way. For example, in my informal studies of inflation (1983a, b), interpretations spring from the stand taken on the issues raised in the present paper.

REFERENCES

Alj, Abderrahmane and Alain, Haurie, "Dynamic Equilibrium of Multigeneration Stochastic Games," *IEEE Transactions on Automatic Control*, February 1983, *A6-28*, 193–203.

Barro, Robert J., "On the Determination of the Public Debt," *Journal of Political Economy*, October 1979, *87*, 940–71.

Basar, Tamer and Olsder, Geert J., *Dynamic Noncooperative Game Theory*, New York: Academic Press, 1982.

Bryant, John and Wallace, Neil, "A Price Discrimination Analysis of Monetary Policy," *Review of Economic Studies*, forthcoming, 1984.

Buiter, Willem H., "The Macroeconomics of Dr. Pangloss: A Critical Survey of the New Classical Macroeconomics," *Economic Journal*, March 1980, *90*, 34–50.

Doan, Thomas, Litterman, Robert and Sims, Christopher, "Forecasting and Conditional Projection Using Realistic Prior Distributions," manuscript, Federal Reserve Bank of Minneapolis, 1983.

Eckstein, Zvi and Eichenbaum, Martin, (1984a) "Inventories and Quantity Constrained Equilibria in Regulated Markets: The U.S. Petroleum Industry, 1947–1972," in T. Sargent, ed., *Energy, Foresight and Strategy*, Resources for the Future, forthcoming 1984.

______ and ______, (1984b) "Oil Supply Disruptions and the Optimal Tariff in a Dynamic Stochastic Equilibrium Model," in T. Sargent, ed., *Energy, Foresight and Strategy*, forthcoming 1984.

Epple, Dennis, Hansen, Lars Peter and Roberts,

Will, "Linear Quadratic Games of Resource Depletion," in T. Sargent, ed., *Energy, Foresight and Strategy*, forthcoming 1984.

Garcia Marquez, Gabriel, *One Hundred Years of Solitude*, New York: Avon Books, 1971.

______, *Chronicle of a Death Foretold*, New York: Alfred A. Knopf, 1983.

Hansen, Lars Peter and Sargent, Thomas J., "Formulating and Estimating Dynamic Linear Rational Expectations Models," *Journal of Economic Dynamics and Control*, February 1980, *2*, 7–46.

______ **and Singleton, Kenneth,** "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, September 1982, *50*, 1269–86.

Keynes, John Maynard, *Collected Writings IV: A Tract on Monetary Reform*, Cambridge: MacMillan, (1923) 1971.

Kydland, Finn and Prescott, Edward C., "Rules Rather Than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, *85*, 473–91.

______ **and** ______, "Time to Build and Aggregate Fluctuations," *Econometrica*, November 1982, *50*, 1345–70.

Litterman, Robert, *Techniques for Forecasting with Vector Autoregressions*, unpublished doctoral dissertation, University of Minnesota, 1980.

______, "A Bayesian Procedure for Forecasting with Vector Autoregressions," Working Paper, Massachusetts Institute of Technology, 1981.

Lucas, Robert E., "Econometric Policy Evaluation: A Critique," in K. Brunner and A. Meltzer, eds., *The Phillips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl. 1976, 19–46.

______ **and Prescott, Edward C.,** "Investment Under Uncertainty," *Econometrica*, September 1971, *39*, 659–81.

______ **and Sargent, Thomas J.,** *Rational Expectations and Econometric Practice*, Minneapolis: University of Minnesota Press, 1981.

______ **and Stokey, Nancy,** "Optimal Fiscal and Monetary Policy in an Economy Without Capital," *Journal of Monetary Economics*, July 1983, *12*, 55–93.

Romer, Paul, "Notes on Existence of Dynamic Competitive Equilibria: Externalities, Increasing Returns and Unbounded Growth," University of Rochester, May 1982.

Sargent, Thomas J., "Interpreting Economic Time Series," *Journal of Political Economy*, April 1981, *89*, 213–48.

______, (1983a) "The Ends of Four Big Inflations," in R. E. Hall, ed., *Inflation: Causes and Effects*, Chicago: University of Chicago Press, 1983.

______, (1983b) "Stopping Moderate Inflations: The Methods of Thatcher and Poincare," in R. Dornbusch and M. Simonsen, eds., *Indexation*, Cambridge: MIT Press, 1983.

______ **and Wallace, Neil,** "Rational Expectations and the Theory of Economic Policy," *Journal of Monetary Economics*, April 1976, *2*, 169–83.

Sims, Christopher A., "Macroeconomics and Reality," *Econometrica*, January 1980, *48*, 1–48.

______, "Policy Analysis with Econometric Models," *Brookings Papers on Economic Activity*, 1:1982, 107–64.

Taylor, John B., "Estimation and Control of a Macroeconomic Model with Rational Expectations," *Econometrica*, September 1979, *47*, 1267–86.

______, "Output and Price Stability: An International Comparison," *Journal of Economic Dynamics and Control*, February 1980, 2, 109–32.

______, "The Swedish Investment Funds System as a Stabilization Policy Rule," *Brookings Papers on Economic Activity*, 1:1982, 57–99.

Thomas, D. M., *The White Hotel*, New York, Pocket Books, 1982.

Townsend, Robert M., (1983a) "Theories of Intermediated Structures," in K. Brunner and A. Meltzer, eds., *Carnegie-Rochester Conference Series on Public Policy: Money, Monetary Policies and Financial Institutions*, 1983, *18*, 221–72.

______, (1983b) "Financial Structure and Economic Activity," *American Economic Review*, December 1983, *73*, 895–911.

Whiteman, Charles, *Linear Rational Expectations Models: A User's Guide*, Minneapolis, University of Minneapolis Press, 1983.

DISCUSSION

ROBERT J. BARRO, University of Chicago: Thomas Sargent's paper covers two issues that are interesting, but I think largely independent. The first is the Lucas critique of policy evaluation, which involves the idea that shifts in the form of the government's behavior imply shifts in the form of the private sector's behavior. The second issue concerns positive vs. normative theories of governmental actions—an issue that has implications for the role of the economist as an adviser for policy.

The Lucas critique applies, for example, to changes in governmental behavior that reflect exogenous events, such as war, natural disasters, and perhaps elections. These developments likely imply shifts in the (conditional) probability distribution for future government policies, which imply shifts in the rational expectations of these policies, which lead to effects on the behavior of the private sector. The same conclusion obtains for shifts in governmental behavior that stem from unobservable causes—or even, unlikely as it seems, from a response of the government's actions to policy advice from economists. (One can think of policy advice as a variable that perhaps influences the actions taken by the public sector.) In each case the crucial matter is the change in the probability distribution for future governmental behavior, combined with the idea that people understand this change. In particular, the Lucas critique applies whether or not there is a role for policy advice. However, if neither advice nor anything else ever shifts the probability distribution for governmental behavior, then the Lucas critique is not so interesting. Also, it seems inconsistent to argue—as Christopher Sims apparently has—that first, the probability distribution for future government behavior does not shift much over short periods, but second, it is conceivable that policy advice could be offered that would lead to a sudden and substantial shift in actual governmental behavior.

Sargent also discusses positive vs. normative theories of government, which is an interesting old issue. My main surprise was in learning that Sims was a major contributor to this literature. (Judging by the disclaimers in Sargent's introduction and fn. 2, I would guess that Sims was equally surprised.) I would have expected references to the substantial work on public choice, including writings by James Buchanan and Gordon Tullock, George Stigler, and Earl Thompson, among others.

It is useful to carry out positive analyses of governmental behavior, possibly viewing the people in the government as attempting to maximize something. Sargent's example about deficits is a good one—the sensible objective of smoothing fluctuations in tax rates over time turns out to have great empirical value in explaining budget deficits. (Having previously been criticized for neglecting Ricardo, I must mention that this idea appears in Pigou and should perhaps be called the Pigovian Theorem.) However, once we think of the government as doing things sensibly—or at least with calculation—then we have to question where policy advice comes in. As economists, we should face that the answer may be nowhere, or at least not much. Possibly there is some role at the "constitutional level" of Buchanan-Tullock, where the choice is about the rules of the game, rather than day-to-day operations. In any event, a positive analysis of what the government actually does is a prerequisite to useful policy advice. For example, we might want to evaluate alternative monetary standards or the draft vs. a volunteer army. In each case we should try first to understand the existing policy—was it motivated by stupidity, or generated by a structure that creates distortions, or was it possibly a reflection of some good reasons that we should know about? In many cases the understanding of actual behavior will give us reasons not to advocate changes in that behavior.

I had more difficulty in appreciating George Perry's paper. Basically, he seems to start from knowledge about how the macroeconomy really operates and about what are useful macro policies. When various theories and developments are evaluated to the extent that they accord with these "facts." For example, micro foundations and rational expectations are okay as long as they give the right answers, which seem mainly to be the ones that Perry had fifteen years ago.

This approach might be satisfactory if we really had a macro model—even with little theoretical base—that worked well empirically. But people did not become dissatisfied with the Keynesian model for no reason. In fact, the main reason was empirical failure, so I cannot accept Perry's statement that the model works well empirically.

As Perry notes, the Keynesian theory has little to say about the determinants of inflation, which is a serious flaw in a model where the behavior of nominal prices and wages is crucial for the determination of quantities. This problem also makes it hard to apply the theory to present-day economies, in which inflation rates are high and volatile.

The Keynesian model also faltered in dealing with supply shocks—because the preoccupation with aggregate demand meant a neglect of supply-side elements. As a related matter, the model generated erroneous predictions about the Phillips curve. For example, there are problems in explaining the rise in inflation during the recessions of 1974–75 and 1980, as well as the fall in inflation during the recovery of 1983. Finally, the Keynesian theory has empirical flaws in its predictions for the effects of monetary and fiscal policies.

According to Perry's paper, the theoretical innovation that will rescue the old macro analysis (presumably including the old advice for activist macro policy) is a model of the "price-tag economy." This setup involves long-lasting implicit contracts between firms and workers, firms and suppliers, firms and customers, and so on. These devices may well be significant for efficiency, but one wonders why we would need the usual Keynesian activist macro policies in such a world. People with these regular relationships do not seem to need the government's help in order to do things efficiently. For example, even if people fix nominal wages for some period (which this line of theory does not explain), they can—as Robert Hall has said—agree to work more when there is more work to do, and vice versa. It is unnecessary in this contractual environment for wages to change from day to day in order to induce the efficient pattern of work and production. In other words, the stickiness of nominal wages and prices is not so important in this type of model. Generally, it is hard to generate significant monetary nonneutrality—or a useful role of activist macro policies—in a fully worked out model with implicit contracts.

ALAN S. BLINDER, Princeton University: Discussants are supposed to be provocative and to disagree with the papers. But Thomas Sargent's paper argues both sides of the issue admirably. And George Perry urges economists to bend their theories to accommodate good sense, not suspend their common sense to accommodate good theory. Since I cannot disagree with either paper, I'll just try to be provocative.

I agree with Perry that rational expectations is a basically sound and important idea which has nothing inherently to do with the belief that labor and product markets are auctions. If prices move sluggishly, it is rational to expect them to do so.

But a curmudgeon would point out the curious nature of rational expectations, given the current state of macroeconomics. "Rational" expectations means model-consistent expectations. But expectations that are consistent with a New Classical model are not consistent with Perry's view of the world, and vice versa. Since no one knows the true model, I think we must at least entertain the notion that people's expectations—though quite rational—are not consistent with the economist's model. And, as Edmund Phelps and others have pointed out, this spells potentially deep trouble for rational expectations.

Perry's main point is that, while the search for microfoundations is critically important, the New Walrasians are barking up the wrong tree. Better micro foundations, he suggests, might be found by taking Arthur Okun's far broader conception of "maximizing behavior" more seriously.

Again, I agree, but must add a pessimistic note. If our job is to explain nominal rigidities, any micro model is going to run into the following difficulty. Since rational, maximizing agents should care only about real wages and relative prices, micro theory will, at best, explain real rigidities, not nominal ones.

So where can we look for an explanation of nominal rigidity? Part of the answer, I think, is that in a decentralized economy with price-setting behavior and heterogeneous goods, no one gets to set his real wage or relative price. The instruments that agents actually control are nominal wages and prices. Or are they? Why don't people enter into indexed contracts? A number of people, including Gray, Fischer, Card and myself, have offered partial answers based on pure neoclassical micro theory. But, to my mind, none of them explains why zero indexing is the norm—except at very high rates of inflation.

My main purpose is to pose this question, not to answer it. But, in the spirit of being provocative, here are two possibilities that would occur only to a desperate man.

Inertia: Current contracts are not indexed because previous contracts were not indexed. And previous contracts were not indexed because inflation was so low that it was not worth incurring the transactions costs necessary to do so. Inertia has no place in the naive neoclassical world that Perry criticizes. But it need not be inconsistent with a broader conception of maximizing behavior such as Okun's—if there is a fixed cost of making a decision (which there surely is).

Money Illusion: Like most economists, I have long rejected money illusion as a form of irrationality. But experience talking with people who are not economists strongly suggests that deflation by a price index is not something that comes naturally. (As Bob Hall once put it, "People can add, but they can't divide.") And even economists have been known to do irrational things. I have probably done at least four irrational things today —including broaching the subject of money illusion on this platform!

I now turn to Sargent's paper, which is an engaging Socratic dialog between Sargent and Sims. I can almost see the two of them strolling through a snowy olive grove in Minneapolis—with Tom trying to put words into Chris's mouth.

The fundamental issue of the paper reminds me of the fun we 11th graders used to have in English class, arguing about whether or not there is free will. To my adolescent mind, it always seemed obvious that there was free will because two people confronted with the same options often make different choices.

But my English teacher, having debated many previous 16-year-olds, pointed out that perhaps each person had a different past history and that, given that history, exercised no effective free will. My English teacher didn't know about budget sets and optimization, or he might have put it this way: each agent has different endowments, and simply makes his optimal choice given these endowments, thereby exercising no free will. Nor did he know about stochastic processes, or he might have added that choices might differ for purely random reasons—even in the absence of "free will."

And I didn't know about utility functions, or I would have objected that an agent's choice depends not only on his endowments, but also on his preferences. Saying that people have different preferences is more or less the same as saying that they have free will.

But, had my high school teacher been trained in economics, he would have pointed out that optimization maps the endowments, choices, and preferences into a decision. Given the same endowments, choices, and preferences, the individual will always make the same decision—except for pure randomness. So there is no free will.

At this point, my naive 16-year-old mouth might have blurted out something like "Preferences can change" or "It's my utility function, and I don't have to maximize it if I don't want to." But, as a grown economist, I'd never say such things.

I have now more or less covered the ground of the Socratic dialog between Sargent and Sims. Briefly, their debate goes like this: if the government optimizes, then its behavior is part of the model, as endogenous as anything else. And how can we talk about a change in the government's policy? If it's already optimal, why would they change it? But, if this is so, then the Lucas critique is nothing to worry about because there are no regime changes. And the program of rational expectations econometrics—to find policy interventions that improve welfare—is futile. So we might as well forecast with simple vector autoregressions and stop giving advice.

Sargent's paper raises a serious question that I would like to address. Is it reasonable to model government policy choices as solutions to individual optimization problems in a fixed (though stochastic) environment? I think not.

First, the government is not an individual. The government thinks with many heads and speaks with many voices; and committees need not have transitive preferences. The outcome of political infighting on something as complex and sweeping as stabilization policy may not lead to the minimum of some quadratic loss function. And changes in the institutional structure of government may change the decisions that emerge. It is not just substance, but also procedure, that matters.

Second, many people find the notion that the government optimizes anything fanciful. Eleven years ago, Steven Goldfeld and I gave a paper suggesting that the U.S. government might have been following an *ad hoc*, but stabilizing, reaction function. Jack Kareken scoffed at the idea of a stable fiscal policy reaction function. Now Sargent, speaking for Sims, sees the government as the continually optimizing dominant player in a differential game. Things have certainly changed in Minnesota! Furthermore, as Sargent mentions, governments change. If we imagine that governments have objective functions, I presume that this function occasionally changes—as when Reagan replaced Carter, or when Volcker replaced Miller. This sounds a little like free will to me.

Finally, the perceived environment is not fixed. I emphasize the word "perceived" because, as suggested earlier, no one knows the true model for sure. We learn and unlearn things all the time. Some of these things change our model of the economy. Hence, even if the government had a fixed objective function that it optimized, its decision rule might change from time to time. For example, the "Keynesian" demand management policies of the postwar period cannot be optimal if you believe in the classical—or New Classical—model. I believe that the federal government's macro model changed between Hoover and Kennedy, and so did policy.

Now, my high school English teacher and I—like Sims and Sargent—could argue interminably about whether this is evidence of "free will" or not. But who cares? It is a reason why the government is not locked in to a fixed pattern of behavior. So I, like Sargent, conclude that we cannot be content with uninterpreted vector autoregressions.

W. D. NORDHAUS, Yale University: Notwithstanding the popular image, economists agree on much. But the key enduring issue of positive economics today continues, as it has for fifty years, to surround the theory and policy concerning business cycles.

We've heard today two polar modern treatments by Perry and Sargent. Sargent argues that we have a well-developed methodology for analyzing dynamic processes (like business cycles); and that we have made substantial progress in understanding such issues and policies. Perry argues that we have a fairly clear idea of how the complex macroeconomy behaves, but a complete microeconomic view eludes us; and the New Classical macroeconomics (*NCM*) provides misleading guidance for policy.

I find myself in sympathy with both positions (as well as with the view expressed by the ghost of Christopher Sims-past). But ultimately, since economics is an empirical science and not a branch of applied mathematics, I find the *NCM* so far has not lived up to its promises. Like the guns of Singapore in 1941, the *NCM* is extremely powerful but trained in the wrong direction. And if macroeconomics is to be reconstructed today, it will need to train the guns of Sargent on the targets of Perry.

The major research error in *NCM* was to use as a paradigm the *efficient markets model*—modernized Walrasian theory. We have heard much criticism of the *market-clearing* assumption in *NCM* models. But we have not yet heard enough, for the *NCM* theorists continue to lean almost completely on this assumption.

Now I would stress, as did Perry, that many central markets don't clear. And it isn't a matter of shrugging it off by saying "they almost clear." They don't come close. For example, studies indicate that prices in auction markets move to erase arbitrage opportunities in a few seconds. Price disequilibria in non-auction-product markets last a month or so. In contractual labor markets, the half-life of a disequilibrium is half a decade.

By analogy, if the speed of adjustment in auction markets is the speed of light, then that of contractual labor markets is 55 miles an hour. You wouldn't think much of a physicist who assumed that the national speed limit is the speed of light; you shouldn't think much of theories that assume markets clear at lightning speed.

At the same time, we know from history of science that empirical observation alone is not persuasive. Heliocentric theories convinced skeptics only after Newton discovered the general theory of gravity. And modern Walrasians will take wage-price stickiness seriously when a sound theoretical basis is constructed. I think Perry is too optimistic in his assessment that there are firm micro foundations for sticky wages and prices. Almost all work on contractual wage and price stickiness concerns real or relative wages and prices. The issue, as I will outline in a minute, concerns nominal wage, price, or contractual rigidity. Robert Barro got it just right in 1972.

In my remaining remarks, I will suggest that there is a general way to understand the phenomena. It surrounds the notion that implicit or explicit contracts are pervasive, are incomplete, are costly to negotiate or renegotiate, and therefore show a great deal of *contract stickiness*.

The background for such a view is the following: a modern economy has become enormously productive because of the profound degree of specialization and interdependence. Most of these interdependencies take place through formal or informal arrangements (or contracts).

Now arrangements are often complex and involve innumerable unwritten details. Important and commonplace examples include the workplace and family life. You will instantly recognize that work and marriage involve more interactions than could even be enumerated.

Moreover, it is in reality costly to negotiate contracts. Implicit marriage contracts take years to negotiate (and years to break). In labor-management bargaining, contracts cost time, managerial attention, and strikes if they fail. Or take OPEC: every time OPEC oil ministers want to change price, they have to expose themselves to terrorist activities and hours of harangues from Libyans and Iranians. To take a final examine, every time GM has a general price change it must gather its 20-person price review committee—hardly a candidate for a clear float. And don't forget that a good contract lawyer costs $200 an hour.

Given these two central features—pervasiveness of cooperative interdependent arrangements and high costs of negotiations—there are important implications:

Contractual Incompleteness. Because of the high cost of contracting, many contracts are incomplete or informal. Labor or marriage contracts cover only a miniscule fraction of life at home or at work. Contracts differentiate between only a tiny number of future states of the world. (Just examine such a contract to see how few state-contingent features there are, and I have yet to see one involving the money supply!)

Contractual Stickiness. Given the high costs of negotiation and renegotiating, people tend to ride out existing contracts even when they are not optimal. (And given the lack of state-contingent features, they are almost always suboptimal.)

Nominal-Price Contracts. A corollary of the first two propositions is that superoptimizing contracts might well be written in nominal (dollar) terms and are *not* 100 percent indexed.

Optimality. The major complaint with such an approach is that it surely is not optimal to have such pervasive contractual stickiness. This is a difficult empirical question, but I suspect long-term (even though incomplete) contracts have considerable utility. No marriage can survive when partners are nightly negotiating about who takes out the garbage. Britain shows what happens when labor-management trust breaks down, while Japan shows how productive an economy can be when bonds of trust are strong.

I think a case can be made that putting such a sticky-contract world into macroeconomics would produce a system tolerably like today's: *Sticky-contract macroeconomics* would (a) show nominal wage and price movements that are inertial but not rigid; (b) produce short-run nonneutrality of money, but long-run neutrality; (c) suggest that the most sticky contracts would occur in those areas (labor markets) where negotiation costs are highest; (d) show why quantity adjustment is the norm in sticky-contract markets —producing cyclical behavior of inventories, employment, layoffs, and quits, quite like that observed; (e) would show highly varied structures over space and time—as contracts gradually but imperfectly adapt to the time and place; (f) would ultimately produce many 100 percent indexed contracts in countries (such as Brazil or Israel) where the variance of general price level movements swamped that of relative prices. In short, *in a world where price contracts are pervasive and costly to negotiate, nominal shocks will have real consequences.*

I would emphasize that Sticky Contract Macroeconomics leaves unanswered (but answerable) one important issue, that of Okun's macroeconomic externality. Having experienced a $1 billion recession, we observe that disinflation—a changing of the inflation-wage norm of existing arrangements—is extremely costly. We do not know whether these enormous macroeconomic costs are just the sum of millions of individual adjustment costs, or whether somehow the sum is much greater than the parts. I suspect that a complete theory along the above lines would find that much of the costs of disinflation are not intrinsic to disinflation. Rather, they are the result of a nonsynchronized adjustment of price and wage inflation norms in which the information about the changing environment is transmitted by inefficient quantity signals rather than by efficient price signals. The answer to this question is enormously important for inflation policy.

Gardner Ackley noted in his 1982 presidential address that, when he was a student, there was no subdiscipline called macroeconomics. Are we heading for a similar fate today, as the carcass of macroeconomics is being carved up and served up to contract theory, search economics, vector autoregression studies, stochastic dynamics, and sociology?

For one, I would hope that this balkanization of macroeconomics will be resisted. Perhaps by training the heavy guns of rational expectations theorists on the right target—non-Walrasian systems—we can reconstruct a useful macroeconomic theory. Failing such an alliance, a future historian may write: "The age of Keynesians and Lucasians is gone. That of supply siders, protectionists, and goldbugs is here. And the glory of economics is extinguished forever."

AMERICAN ECONOMIC ASSOCIATION

PROCEEDINGS OF THE NINETY-SIXTH ANNUAL MEETING

SAN FRANCISCO, CALIFORNIA
DECEMBER 28–30, 1983

THE JOHN BATES CLARK AWARD

Citation on the Occasion of the Presentation of the Medal to

JAMES HECKMAN

December 29, 1983

James Heckman's research has changed the face of labor economics, econometrics, and demography. His work on panel data and selection problems has set the standard for analysis of microeconomic cross-section, time-series data. His treatment of dynamic models, clarifying the observable implications of heterogeneity and state dependence, has advanced our understanding of economic phenomena regarding durations, particularly the duration of unemployment. The technique he originally proposed for handling the selection problem in analyzing wage rates of women, adding an estimated hazard rate term to compensate for the nonzero conditional mean of the disturbance in the selected data, has become so universal that econometrics students are routinely taught how to "heckit" regression equations. His recent work on nonparametric problems associated with the analysis of longitudinal data is on the frontier of statistics, and has deep implications for econometrics. His innovations promise to be of lasting importance in economics and across the social sciences.

Minutes of the Annual Meeting
San Francisco, California
December 29, 1983

The Ninety-Sixth Annual Meeting of the American Economic Association was called to order by President W. Arthur Lewis at 9:28 P.M., December 29, 1983, in the Continental Ballroom of the San Francisco Hilton Hotel. The minutes of the meeting of December 29, 1982 were approved as published in the *American Economic Review, Papers and Proceedings*, May 1983, pages 385–86.

The Secretary (C. Elton Hinshaw), Treasurer (Rendigs Fels), Managing Editor of the *Journal of Economic Literature* (Moses Abramovitz), Managing Editor of the *American Economic Review* (Robert Clower), and the Director of *Job Openings for Economists* (Hinshaw) discussed their written reports which were distributed at the meeting. (See their reports published elsewhere in this issue.)

The Treasurer reported that real dues had fallen about 25 percent since 1976, but the financial outlook of the Association is bright. The 1984 budget projects a surplus and it is anticipated that net worth will continue to grow relative to expenditures. He announced that nominal dues will be held constant for the next few years and that the Executive Committee would welcome suggestions for worthwhile new programs the Association might undertake. William Vickrey, from the floor, expressed his disagreement with the Association's policy of discriminating against libraries in the dues structure. The Association is in good financial condition; libraries have significant financial problems.

Abramovitz noted in his report that Naomi Perlman, Associate Editor of the *Journal of Economic Literature* had resigned as of May 31, 1984, because she and the Association had failed to reach an agreement on the arrangements needed to continue the bibliographic work in Pittsburgh. He expressed a lively appreciation for her past work and looked forward to an early opportunity to recognize her past contributions. Arrangements had been made with Drucilla Ekwurzel and Asatoshi Maeshiro to continue doing the bibliographic work of the *JEL* in Pittsburgh. Both have worked in that office for several years.

The Secretary presented the following resolution, which was adopted unanimously by voice vote:

> BE IT RESOLVED that the American Economic Association express its appreciation to the members of the 1983 Allied Social Science Associations' Convention Committee, chaired by Michael Keran and Hang-Sheng Cheng (both of the Federal Reserve Bank of San Francisco), for their cooperation and hard work. With the Committee's help, the tasks of organizing and running the meetings have been made much less onerous.

Lewis then introduced Charles L. Schultze, the President of the Association for 1984, to the assemblage.

The meeting was adjourned.

Respectfully submitted,
C. ELTON HINSHAW, *Secretary*

Minutes of the Executive Committee Meetings

Minutes of the Meeting of the Executive Committee in Washington, D.C., March 18, 1983.

The first meeting of the 1983 Executive Committee was called to order at 10:15 A.M. on March 18, 1983 in the Caucus Room of the Washington Hilton Hotel, Washington, D.C. Members present were W. Arthur Lewis (presiding), Gardner Ackley, Elizabeth E. Bailey, William Baumol, Robert W. Clower, Rendigs Fels, Ann F. Friedlaender, Robert J. Gordon, C. Elton Hinshaw, Juanita M. Kreps, Edmund S. Phelps, Charles L. Schultze, A. Michael Spence, and Joseph E. Stiglitz. Leo Raskind, Counsel, was also present. Present for parts of the meeting were members of the Nominating Committee (Moses Abramovitz, Ralph d'Arge, T. Aldrich Finegan, John G. Gurley, Alfred E. Osborne, Jr., and Anita A. Summers), members of the Honors and Awards Committee (Robert Eisner, Daniel McFadden, Dale T. Mortensen, and William Vickrey) and Naomi Perlman, Associate Editor of the *Journal of Economic Literature*.

Minutes. The minutes of the meeting of December 27, 1982 were corrected for typographical errors and then approved.

Report of the Secretary (Hinshaw). The Secretary reported that the 1983 annual meetings will be held in San Francisco, December 28–30. The schedule for subsequent meetings is Dallas (1984), New York (1985), and New Orleans (1986). Boston, New York, Chicago, and Washington are being considered as possible sites for the 1987 meetings.

In response to a request from Amnesty International to "adopt" a prisoner of conscience, Dr. Marko Veselica, the Secretary, as instructed, wrote Amnesty International to acquire more information about the case. He wrote that the Association limits its involvement in such cases to those that meet three criteria: the person is an economist, academic freedom is at issue, and there is substantial hope that AEA help will be effective. He asked if Dr. Veselica's case met the three criteria and if there were other cases. Amnesty International, in the person of Pierre Deguise, responded that the Veselica case meets the criteria but did not respond about other cases. The Executive Committee directed the Secretary to inquire again about other cases that might be more urgent and to find out if Deguise could speak on behalf of the organization.

Effective February 1, 1983, AEA members are entitled to a 15 percent discount on "Standard Unlimited Mileage Rates" at all Hertz corporate and participating licensee locations in the United States. This discount applies to basic or standard rates only. The old discount rate was 5 percent.

Because several members of the Executive Committee are funding travel to the December meeting out of their own pockets or on research grants, it was VOTED to change the current policy of not reimbursing members for travel to one that pays for travel and two nights lodging.

Report of the Treasurer (Fels). The budget for 1983, as approved by the Executive Committee on December 27, 1982, projects a surplus of $133 thousand. The surplus for 1982 was $454 thousand. The net worth of the Association is more than adequate. Since costs are rising faster than revenues, the surpluses cannot be expected to continue indefinitely. It was VOTED to raise dues $3\frac{1}{8}$ percent effective January 1, 1984, continuing the policy of raising dues less than the inflation rate. It was also VOTED to appoint an *ad hoc* committee to consider options available to the Association for dealing with past and projected surplus; the committee is to organize a "menu of options" and present it to the Executive Committee.

The Treasurer reported that the present rate structure does not conform to Postal Service requirements for second class mailings. The requirements can be satisfied by establishing a subscription rate for individuals of $65. The adverse financial effect will be minimal, since only a few of our subscribers, who pay $100, are individuals. It was VOTED to establish a subscription rate for individuals.

Report of the Editor of the American Economic Review (Clower). Clower reported that the number of submissions continues to increase (perhaps partly due to quick rejections) and the *Proceedings* volume of the journal is in press and will be out nicely on time. He stated that he would like to end his term by the end of 1985 (or earlier), rather than 1986. It was decided to appoint a search committee to seek his replacement.

Report of the Pittsburgh Office of the Journal of Economic Literature (Perlman). Perlman, Associate Editor of the *JEL*, reported that the bibliographic data base had been on-line with Dialog since February 4, 1983. Royalties from its use should amount to about $1,000. She has received inquiries from some libraries that wish to lease the data base and will continue to explore these possibilities. The 1978 *Index* will appear this May. She is dissatisfied with the current distributor of the *Index*, Richard D. Irwin, Inc., and wishes to consider moving to another distributor. The Cambridge University Press was mentioned as a possibility. Upon her recommendation, it was VOTED to join the National Federation of Information and Abstracting Services.

1983 Program (Schultze). Schultze, the 1983 Program Chairman, announced that Thomas Schelling had been invited to give the Ely Lecture and that a special session commemorating the 100th anniversary of the death of Marx and the births of Keynes and Schumpeter had been organized. Giersch, Gurley, and Patinkin would present the papers for the session. At present, about 25 invited sessions, about 20 self-generated sessions, about 20 contributed ones, and about 20 that are cosponsored have been arranged. After a general discussion of the appropriate nature of the *Proceedings* issue of the *AER* and the availability, or lack of it, of the issue to relative unknown economists as a publishing outlet, it was decided that William Baumol would discuss with Lawrence Klein the procedure Klein used in selecting papers for publication in the *Proceedings* when Klein was President-elect and report back to the Executive Committee. Klein had selected several contributed papers for publication after they had been presented at the annual meetings.

Honors and Awards (McFadden). Acting together as an electoral college, the Committee on Honors and Awards and the Executive Committee VOTED to award the John Bates Clark Medal to James J. Heckman.

Nominating Committee (Abramovitz). The Electoral College, consisting of the Nominating and Executive Committees meeting together, chose Charles P. Kindleberger as the nominee for President-elect, and Abram Bergson and James M. Buchanan as Distinguished Fellows. Abramovitz reported the following nominees for other offices: for Vice-President (two to be chosen), Zvi Griliches, Albert O. Hirschman, Hollis B. Chenery, and Robert L. Heilbroner; for members of the Executive Committee (two to be chosen), Michael R. Darby, Stanley Lebergott, Janet L. Norwood, and Victor R. Fuchs.

Other Business. It was VOTED to elect the Secretary and the Treasurer to additional three-year terms. Their terms will now end December 31, 1987. It was decided that each President, upon the expiration of his or her term on the Executive Committee, would write a report reflecting on the problems and successes during his or her four years on the Executive Committee. The Secretary was instructed to develop orientation materials for incoming members of the Committee and to attempt to give more guidance to the Committee concerning future problems and prospects. The President was asked to set in motion whatever steps are necessary to amend the bylaws to allow noneconomists to be elected Distinguished Fellows.

The meeting adjourned at 3:45 P.M.

Minutes of the Meeting of the Executive Committee in San Francisco, California, December 27, 1983.

The second meeting of the 1983 Executive Committee was called to order at 10:03 A.M. on December 27, 1983 in the Continental Parlor I of the San Francisco Hilton Hotel, San Francisco, California. Members present were W. Arthur Lewis (presiding), Gardner Ackley, Elizabeth E. Bailey, William Baumol, Robert W. Clower, Rendigs Fels, Ann F. Friedlaender, Robert J. Gordon, C. Elton Hinshaw, Juanita M. Kreps, William D. Nordhaus, Edmund S. Phelps, Charles L. Schultze, A. Michael Spence, and Joseph E. Stiglitz. Leo Raskind, Counsel of the Association, and A. W. Coats were also present. Charles P. Kindleberger, Victor R. Fuchs, and Janet L. Norwood, newly elected officers for 1984, were present as guests. Present for parts of the meeting were Marcus Alexis, Kenneth J. Arrow, Barbara R. Bergmann, Donald J. Brown, John H. Cumberland, W. Lee Hansen, Roberta Miller, Naomi Perlman, and Lloyd G. Reynolds.

President Lewis welcomed Kindleberger, Fuchs, and Norwood (new members of the 1984 Executive Committee) to the meeting and thanked those whose terms were expiring (Baumol, Kreps, Phelps, Bailey, and Gordon) for their services to the Association. Noting that the Executive Committee had delegated to him the authority and responsibility for attempting to resolve the issues raised by Naomi Perlman, Associate Editor of the *Journal of Economic Literature*, concerning the operation and status of the bibliographic sections of the *JEL*, Lewis reported that he had met with Perlman in Pittsburgh to review the situation with her. After several months of negotiation, which included hiring a consultant to review the salary structure of the Pittsburgh and the *American Economic Review* staffs, Lewis received a letter from Perlman's lawyer advising him that Perlman was giving six months notice that she would resign effective May 31, 1984 if certain conditions were not met. Lewis consulted both individually and collectively with Abramovitz (Editor of the *JEL*), Ackley (past-President), Schultze (President-elect), Hinshaw (Secretary), Fels (Treasurer), and Raskind (Counsel). It was decided to accept her resignation. Arrangements are being made to continue the bibliographic functions in Pittsburgh under the direction of Drucilla Ekwurzel, currently an Assistant Editor, and Asatoshi Maeshiro of the University of Pittsburgh. Both have worked in the Pittsburgh office for several years. Following the discussion of Lewis' report it was decided to reactivate the committee originally appointed to examine the long-term arrangements needed for the journal, including its organizational structure.

Minutes. The minutes of the meeting of March 18, 1983 were approved as written and circulated.

Consortium of Social Science Associations (Miller). Miller, the Executive Director of the Consortium (COSSA) briefly summarized the annual report which had been sent to members of the Executive Committee. COSSA attempts to represent the interests of the social and behavioral science research community. It focuses on obtaining support for the social science research budgets in the National Science Foundation and other research agencies. A copy of the full report on its 1983 activities can be obtained by writing COSSA, 1775 Massachusetts Avenue, N.W., Suite 300, Washington, D.C. 20036. The AEA contributes $35,000 per year to the support of COSSA's activities. It was decided to continue support at that level for 1984.

Report of the Managing Editor of the Journal of Economic Literature (Abramovitz). Abramovitz summarized his written report (see elsewhere in this issue). In addition, he discussed the prospects for a smooth transition of operations in the Pittsburgh office from Perlman to Ekwurzel and Maeshiro, and thanked Perlman for her work during the past fourteen years.

He is undertaking a systematic review of the journals indexed by the *JEL* by establishing panels to review the list by fields and languages. The current list is large and of uneven quality. He is also attempting to expand and systematize the pool of book reviewers available; he particularly wants to include younger members of the profession in the pool.

Acting on his recommendation, the Executive Committee VOTED to approve the appointment of Carolyn Shaw Bell, Robert Eisner, Duncan Foley, Victor Fuchs, Jack Hirshleifer, and Roger Noll to the Board of Editors.

Perlman, Associate Editor of the *JEL*, announced her resignation effective May 31,

1984. She also reported that the bibliographic data base had been on-line with Dialog since February 1983, and the 1978 *Index of Economic Articles* was published during the past year.

Report of the Secretary (Hinshaw). The Secretary announced the schedule for subsequent meetings: Dallas in 1984, New York in 1985, and New Orleans in 1986. It was agreed that either Boston or Chicago would be selected as the site for 1987. He noted that on the 1978 and 1982 preference ballots, members had ranked "Immediately after New Year's" second and first (respectively) as the time to hold the annual meetings. It was VOTED to continue to hold the meetings on December 28–30.

The Secretary had again inquired of Amnesty International about cases that might be more urgent than Dr. Marko Veselica's. The organization's response was not responsive. It was agreed that the President would call Amnesty International to seek more information about Veselica's case and others. If, in his judgment, a basis exists for action, the Secretary would circulate a letter to members of the Executive Committee on Veselica's behalf. Members who wanted to sign would do so as individuals, not on behalf of the Executive Committee. That is, the Executive Committee as such would not act collectively on behalf of the Association.

At its March 1983 meeting, the Executive Committee asked the President to set in motion whatever steps were necessary to allow noneconomists to be elected Distinguished Fellows of the Association. It was decided that if the Nominating Committee and the Executive Committee, acting together as an Electoral College, decide that a person merits being elected Distinguished Fellow, there is a presumption that the person is probably an economist, at least some of the time. No action was needed.

The Secretary called attention to correspondence from Joseph Hasson concerning a code of ethics for economists. As in the past when the issue has been considered, it was decided to take no action.

Report of the Managing Editor of the American Economic Review (Clower). Clower briefly reviewed his written report (see elsewhere in this issue). He indicated that he was eager for the Search Committee to find his replacement; the sooner the better. It was VOTED to approve his nominees to the Board of Editors: Clive D. Bull, Michael R. Darby, Philip E. Graves, Meir Kohn, Susan Woodward, and Leslie Young.

Report of the Treasurer (Fels). The Treasurer projected an operating deficit of $62,000 and income from investments of $360,000. On balance, the expected surplus for 1984 is $298,000. It was VOTED to adopt the budget as submitted. See his full report and the 1984 budget elsewhere in this issue.

Report of the Director of Job Openings for Economists (Hinshaw). See his written report elsewhere in this issue.

Committee on Economic Education (Hansen). Hansen, Chairman of the Committee, reviewed its activities. The Committee was established to help improve the teaching of economics in colleges, universities and, with the cooperation of the Joint Council on Economic Education, high schools. Major activities during the past years have been the development of a teacher training program (some 20 to 25 graduate programs have adopted it), a major research project on the characteristics of undergraduate economics majors and programs, and (in cooperation with the Joint Council) the establishment of the *Journal of Economic Education*. Plans for the future include an analysis of the use of computers in teaching economics, a systematic review of learning theory and how it can be used to improve the structure of undergraduate courses, and a research project on the differences between males and females in their performances in economics courses.

Committee on the Status of Minorities in the Economics Profession (Alexis). Alexis reported that the Summer Program had been transferred from Yale to Wisconsin and that it would remain there through 1985. The 1983 program had 26 students participate (16 males, 10 females). It was estimated that one-third of the students would be successful in a good graduate program. He noted that the past summer faculty did not have a member of a minority group and expressed some concern about the lack.

He announced an additional fellowship program—doctoral dissertation grants for third- and fourth-year students. The program is a joint effort of the Consortium for Minority Graduate Study in Economics and the Federal Reserve System. The program will provide opportunities for summer research internships at the Board of Governors and Federal Reserve District Banks in addition to the support of campus-based research.

The Consortium for Minority Graduate Study in Economics was being reorganized. This group, started in 1981, consisted of Yale, M.I.T., Michigan, Northwestern, and Stanford. The Consortium had been willing to support ten fellowships per year for entering minority students. The number of candidates nominated has been disappointing. The Consortium Steering Committee has agreed to expand the Consortium and to make fellowships available to Ph.D. candidates at any institution provided certain conditions are met. Inquires and nominations should be sent to Marcus Alexis, Department of Economics, Northwestern University, Arthur Andersen Hall, 2003 Sheridan Road, Evanston, IL 60201.

Alexis anticipates having an unspent balance of $150,000 from the original $350,000 Rockefeller grant when it expires June, 1984. He is applying for an extension.

In concluding, he noted that the AEA had supported the activities of the Committee since 1970 with an appropriation of $10,000 each year. At the beginning, this represented about 10 percent of the budget. It now represents about 3 percent. He advocated a substantial increase in the level of commitment, perhaps to $25,000. But he did not ask for an addition to the budget for 1984.

Report of the Representative to the Council of Professional Associations on Federal Statistics (Cumberland). Cumberland briefly reviewed the report that is printed elsewhere in this issue. It was VOTED to increase the AEA contribution to $10,000. The discussion raised questions about the role of COPAFS, its effectiveness, and the larger responsibilities of the AEA concerning the quality of federal statistics. Cumberland and Gary Fromm were asked to write a detailed report for the March meeting; the Secretary was instructed to invite Katherine Wallman, Executive Director of COPAFS, to speak at that meeting.

International Economic Association (Arrow). Arrow, AEA representative on the International Economic Association Council and newly elected President of the IEA, reported on some of the activities of that organization. The IEA is a federation of national economic associations (57 countries are represented). It organizes two or three international conferences each year, and a triannual international congress of economists. This year the congress is held in Madrid, Spain. India is the leading candidate for the host country to the next congress in 1985. The current bylaws of the IEA provide that three of the founding countries (U.S., U.K., and France) would have two votes each on the Council. All other countries have one. The IEA is considering changing its bylaws to provide for only one vote per member association. It was VOTED to approve the change if it is made.

Committee on the Status of Women in the Economic Profession (Bergmann). Bergmann reviewed her written report (published elsewhere in this issue) and raised three issues of concern: the employment of women, especially at the top positions in the profession; the need for blind refereeing for the *American Economic Review*; and sex bias in textbooks. A lively discussion ensued. No specific action was requested; no action was taken.

Committee on U.S.–Soviet Exchange (Reynolds). His written report is available elsewhere in this issue.

Surplus Committee (Schultze). Schultze reported that the Committee agreed that there is no reason to accumulate surpluses beyond a safe and prudent provision for contingencies. A ratio of net worth to expenditures of about 0.8 to 1.0, especially if based on conservative assumption, would be adequate. The Committee agreed that if, with this provision for contingencies, uses of the surplus could be found in the nature of *public goods* for the profession and the membership, they should be given precedence over further reductions in real dues. They agreed that there were, at least tentatively, two potential uses of the funds that might be preferred to dues reductions.

First, the Committee recommended that the Executive Committee commission a study of the possibility of founding a new nontechnical journal of economics and underwrite the initial expenses of the investigation. The objective of the new journal would be to make available, in relatively nontechnical form, the results of recent theoretical and applied research. It would be aimed at a variety of audiences: the large number of the profession with heavy teaching loads or who work full time at government or business positions and are unable to keep up with the latest developments as they are typically made available; members of other professions (lawyers, engineers, political scientists, and journalists) who need to keep abreast of economics; and interested, thoughtful laypersons. It was VOTED to empower the President, in consultation with the Surplus Committee and other officers of the Association, to hire someone to conduct the feasibility study and to establish a Steering Committee for the project. The Steering Committee would report back on the results of the study, including establishing a schedule for the project and budget requirements.

The second project recommended was to devote about $50,000 each year to a portfolio of projects essentially dealing with the American economics profession as it relates to the rest of the world. Several possibilities were suggested but none were received with general approbation. It was agreed that the Surplus Committee would report again at the March meeting. It was also agreed that (1) the surplus should not be reduced by a significant decrease in dues, but nominal dues should not be raised (real dues should fall at the rate of inflation), and (2) the Association should not continue to accumulate surpluses beyond a prudent and safe level.

Report on the 1984 Program (Kindleberger). The program will emphasize international economics and economic history. Gerard Debreu, Nobel Laureate, will be honored at a luncheon, Alexander Lamfalussy will be the AEA-AFA luncheon speaker, and Sir Alec Cairncross will deliver the Ely Lecture. Some 20-plus sessions are taking shape, and he anticipates organizing 15 to 20 more. He has insisted that sessions should consist of no more than three papers and two discussants. Published papers will be footnoted to indicate who the discussants were and stating that copies of the discussion can be sought of them by letter. He has also urged that session organizers seek new faces for their panels.

Report of the Committee on the Centennial (Borts). On behalf of the Committee, George Borts submitted a written report. The Committee recommended that a program of events for a day of celebration be appended to the December 1985 meetings. The recommended program would include four lecturers on the history of American economics and of American economic thought; five roundtables on current and future issues of economic analysis, public policy, and professional activity; and an evening dinner meeting. This last event should invite the attendance of all present and past officers, fellows, and honorees of the Association. These recommendations met with the approval of the Executive Committee.

The meeting was adjourned at 5:30 P.M.

C. ELTON HINSHAW, *Secretary*

Report of the Secretary for 1983

Annual Meetings. In 1984 the annual meeting will be held in Dallas, Texas on December 28–30. The schedule for subsequent meetings is December 28–30, 1985 in New York and 1986 in New Orleans. Employment services will be provided at these annual meetings beginning December 27.

National Registry. The National Registry for Economists continues to be operated on a year-round basis by the Illinois State Employment Service. Economists looking for jobs and employers are urged to register. This is a placement service that maintains the anonymity of employers. The Association is indebted to the Registry for assistance and supervision of the employment service provided at the annual meetings. Employers are reminded of the Association's bimonthly publication, *Job Openings for Economists,* and their professional obligation to list their openings.

Membership. The total number of members and subscribers is shown in Table 1. The total has fluctuated between 26,000 and 26,500 since 1975, when it reached an all time high of 26,787. Since then the increase in memberships has offset the decline in subscribers.

TABLE 1—MEMBERS AND SUBSCRIBERS
(End of Year)

	1981	1982	1983
Class of Membership			
Annual	16,738	16,771	16,728
Junior	1,800	1,895	1,998
Life	385	384	383
Honorary	32	32	31
Family	374	397	423
Complimentary	607	607	599
Total Members	19,936	20,086	20,162
Subscribers	6,291	5,171	5,986
Total Members and Subscribers	26,227	26,257	26,148

Permission to Reprint and Translate. Official permission to quote from, reprint, or translate and reprint articles from the *American Economic Review* and the *Journal of Economic Literature* totaled 176 in 1983 compared to 230 in 1982. Upon receipt of a request for permission to reprint an article, the publisher or editor making the request is instructed to get the author's permission in writing and send a copy to the Secretary as a condition for official permission. The Association suggests that authors charge a fee of $150, but they may charge some other amount, enter into a royalty arrangement, waive the fee, or refuse permission altogether.

Elections. In accordance with the bylaws on election procedures, I hereby certify the results of the recent balloting and report the actions of the Nominating Committee and the Electoral College.

The Nominating Committee, consisting of Moses Abramovitz, Chair, Ralph d'Arge, T. Aldrich Finegan, John G. Gurley, Daniel J. B. Mitchell, Alfred E. Osborne, Jr., and Anita A. Summers submitted the nominations for Vice-Presidents and members of the Executive Committee. The Electoral College, consisting of the Nominating Committee and Executive Committee meeting together, selected the nominee for President-elect. No petitions were received nominating additional candidates.

President-Elect
Charles P. Kindleberger

Vice President	*Executive Committee*
Hollis B. Chenery	Michael R. Darby
Zvi Griliches	Victor R. Fuchs
Robert L. Heilbroner	Stanley Lebergott
Albert O. Hirschman	Janet L. Norwood

The Secretary prepared biographical sketches of the candidates and distributed ballots last summer. On the basis of the canvass of the ballots, I certify that the following persons have been duly elected to the

respective offices:
President-elect (for a term of one year)
Charles P. Kindleberger
Vice-Presidents (for a term of one year)
Zvi Griliches
Robert L. Helibroner
Executive Committee (for a term of three years)
Victor R. Fuchs
Janet L. Norwood

Number of legal ballots	5,488
Number of invalid envelopes	179
Number of envelopes received after October 1	40
Number of envelopes returned	5,707

AEA Staff. Mary Winer, Administrative Director, Kimberly Adair, Norma Ayres, Ersye Burns, Marcia McGee, Violet Sikes, Dale Wagner, and Jacquelyn Woods handle the day-to-day operations of the Association. I wish to thank them for their efficient and dedicated work. Without them little would get accomplished.

Committees and Representatives. Listed below are those who served the Association during 1983 as members of committees or representatives. The year in parentheses indicates the final year of the term to which they have been appointed. On behalf of the Association, I wish to thank them all for their services.

Ad Hoc Committee on Financial Reporting and Changing Prices
Franco Modigliani, *Chair*
Charles Christenson
Robert Kaplan
Richard W. Leftwich
G. William Schwert
John B. Shoven

Ad Hoc Committee on Relations with IEA
Anne O. Krueger
Franco Modigliani
Paul A. Samuelson
C. Elton Hinshaw

Budget Committee
Rendigs Fels, *Chair*
Elizabeth E. Bailey (1983)
Ann F. Friedlaender (1984)
A. Michael Spence (1985)
W. Arthur Lewis, *ex officio*
Charles L. Schultze, *ex officio*

Census Advisory Committee
Rosanne Cole, *Chair* (1983)
Marcus Alexis (1983)
Victor R. Fuchs (1983)
Zvi Griliches (1983)
Sherwin Rosen (1983)
Norman J. Simler, (1983)
Morris A. Adelman (1984)
Martin H. David (1984)
Sidney L. Jones (1984)
Edwin Mansfield (1984)
Lawrence Chimerine (1985)
Ronald L. Oaxaca (1985)
Joel Popkin (1985)
Richard Quandt (1985)
Ann D. Witte (1985)

Committee on Economic Education
W. Lee Hansen, *Chair* (1984)
George L. Bach (1983)
Marianne A. Ferber (1983)
Herbert Stein (1983)
Michael K. Salemi (1984)
John J. Siegfried (1984)
Kalman Goldberg (1985)
Campbell R. McConnell (1985)
Rendigs Fels, *ex officio*

Economics Institute Policy and Advisory Board
Edwin S. Mills, *Chair* (1986)
Bent Hansen (1984)
Louis T. Wells (1984)
Robert E. Evenson (1985)
W. Lee Hansen (1985)
John R. Moroney (1986)
Dwight H. Perkins (1987)
G. Edward Schuh (1987)

Finance Committee
Rendigs Fels, *Chair*
Robert J. Genetski (1983)
Sidney Davidson (1984)
Robert Eisner (1985)

Committee on Honorary Members
Richard A. Musgrave, *Chair* (1986)
Hendrik S. Houthakker (1984)
George J. Stigler (1984)
Hal Varian (1986)
Richard E. Caves (1988)
Franco Modigliani (1988)

Committee on Honors and Awards
Anne O. Krueger, *Chair* (1983)
Dale T. Mortensen (1983)
Daniel McFadden (1985)
Oliver E. Williamson (1985)
Robert Eisner (1987)
William Vickrey (1987)

Nominating Committee
Moses Abramovitz, *Chair*
Ralph d'Arge
T. Aldrich Finegan
John G. Gurley
Daniel J. B. Mitchell
Alfred E. Osborne, Jr.
Anita A. Summers

Committee on Political Discrimination
Martin Bronfenbrenner, *Chair* (1985)
Herbert Gintis (1983)
Richard R. Nelson (1983)
Harold C. Barnett (1984)
Anne P. Carter (1984)
Lester C. Thurow (1985)

Search Committee for Editor of American Economic Review
Lawrence R. Klein, *Chair*
Robert Eisner
Claudia Goldin
Robert H. Haveman
Robert L. Heilbroner
Joseph A. Pechman
Joseph E. Stiglitz

Committee on the Status of Minority Groups in the Economics Profession
Marcus Alexis, *Chair* (1983)
Donald J. Brown, Chair (1986)
Jeffrey G. Williamson (1983)
Bernard Anderson (1985)

Committee on the Status of Women in the Economics Profession
Barbara R. Bergmann, *Chair* (1985)
Irma Adelman (1983)
Janet C. Goulet (1983)
Jean A. Shackelford (1983)
Monique Garrity (1984)
Joan G. Haworth (1984)
Nancy D. Ruggles (1984)
Gail Wilensky (1984)
Joseph A. Pechman (1985)
Cordelia W. Reimers (1985)
Aleta A. Styers (1985)
W. Arthur Lewis, *ex officio*

Surplus Committee
Charles L. Schultze, *Chair*
Gardner Ackley
Elizabeth E. Bailey
Rendigs Fels
Ann F. Friedlaender
Juanita Kreps
Joseph E. Stiglitz
A. Michael Spence
W. Arthur Lewis, *ex officio*

Committee on U.S.–Soviet Exchanges
Lloyd G. Reynolds, *Chair* (1983)
Abram Bergson (1983)
Joseph A. Pechman (1983)
Richard N. Rosett (1983)

Council and Other Representatives

American Association for the Advancement of Science Section K on Social, Economic, and Political Sciences
Roger Bolton (1983)

American Association for the Advancement of Slavic Studies
Judith Thornton (1985)

AEA/SSRC–Joint Committee on U.S.–China Exchanges
Gregory Chow, *Co-Chair* (1983)
Kenneth J. Arrow
Lawrence R. Klein
Theodore W. Schultz

American Council of Learned Societies
C. Elton Hinshaw (1986)

Review Board of the American Statistical Association–Bureau of the Census Fellowships
Zvi Griliches

Consortium of Social Science Associations (COSSA)
Henry J. Aaron
Joseph A. Pechman

Council of Professional Associations on Federal Statistics (COPAFS)
John H. Cumberland (1983)
Gary Fromm (1983)
Edward F. Denison (1984)

Eighth Symposium on Statistics and Environment–Steering Committee
Paul Portney (1984)

Federal Statistics Users Conference
Paul Wonnacott (1985)

Internal Revenue Service Conference–Tax Administration Research Strategies
Harvey Galper (1985)

International Economic Association
Kenneth Arrow (1984)
C. Elton Hinshaw (1985)

Policy Board of the Journal of Consumer Research
Louis L. Wilde (1985)

National Archives Advisory Council–General Services Administration
William N. Parker (1983)

National Bureau of Economic Research
Carl F. Christ (1984)

Social Science Research Council and SSRC Committee on Programs and Policy
Hugh Patrick (1984)

U.S. National Commission for UNESCO
Walter S. Salant (1985)

Representatives of the Association on Various Occasions—1983

Inaugurations
James F. Watkins, Gainesville Junior College
Kenneth Jones

ASSA 1983 Convention Committee

Michael Keran, *Chair*
Hang-Sheng Cheng, *Vice Chair*
Barbara Weaver, *Convention Manager*
Steve Teigland
William D. Hermann
Don Chaffee
Walter Yep
Ronald Supinski
Violet Sikes
Bent Hansen
Janet Aschenbrenner
Marlene Hall
Thomas Thomson
Norma Ayres
William Wade

C. Elton Hinshaw, *Secretary*

Report of the Treasurer For the Year Ending December 31,1983

The finances of the American Economic Association are robust. Since the price of subscriptions to libraries and other institutions was raised to $100 in 1981, the Association has had substantial surpluses, which have been augmented in the last two years by capital gains from the rise in the stock market.

A special committee has been appointed by the President to consider what policy the Association should pursue with respect to the surpluses. I was asked to project revenues and expenses for the five-year period 1983–87 to help the Surplus Committee in its deliberations. With assistance from my colleague, William W. Damon, I prepared six different projections using different assumptions. The assumptions of the main variant include: inflation of 4.5 percent per year throughout 1983–87; a real total rate of return on equi-

TABLE 1—AMERICAN ECONOMIC ASSOCIATION BUDGETS, 1983–84
(Thousands of dollars)

	First Nine Months Actual (Unaudited)		Full Year		
			Actual	Budgeted	
	1982	1983	1982	1983	1984[a]
REVENUES FROM DUES AND ACTIVITIES					
Membership dues	$ 540	$ 564	$ 729	$	$
Nonmembership subscriptions	479	471	645	[b]	[b]
Subtotal	1,019	1,035	1,375	1,385	1,425
Job Openings for Economists subscriptions	16	18	25	30	30
Advertising	62	69	87	95	100
Sales of *Index of Economic Articles*	43	9	64	50	50
Sales of copies, etc.	26	21	32	35	28
Sale of mailing list	23	26	38	40	42
Annual meeting	37	34	37	15	25
Sundry	34	42	41	40	50
Total Operating Revenue	**1,260**	**1,255**	**1,700**	**1,690**	**1,750**
PUBLICATION EXPENSES					
American Economic Review	349	370	452	493	495
Journal of Economic Literature	458	495	626	708	755
Directory Publication	41	45	55	75	65
Job Openings for Economists	27	34	46	50	60
Index of Economic Articles	19	5	29	30	26
Subtotal	895	948	1,207	1,356	1,401
OPERATING AND ADMINISTRATIVE EXPENSES					
General and administrative	198	196	268	287	311
Committees	26	31	49	50	60
Consortium of Social Science Associations, etc.	48	35	48[c]	35	45[d]
Federal income taxes	4	–	3	5	5
Subtotal	270	263	368	377	421
Total Expenses	**1,165**	**1,212**	**1,575**	**1,733**	**1,822**
OPERATING INCOME (LOSS)	95	43	124	(43)	(72)
INVESTMENT GAINS	104	164	330	170	360
SURPLUS (DEFICIT)	$ 199	$ 207	$ 454	$ 127	$ 288

[a] As approved by the Executive Committee on December 27, 1983.
[b] Revenues for dues and subscriptions are not projected separately.
[c] Includes $10,000 for Economics Institute; $3,000 for COPAFS; $35,000 for COSSA.
[d] Includes $10,000 for COPAFS and $35,000 for COSSA.

ties of 5.5 percent per year; real increases of salaries and wages of 2 percent per year; and no change in (nominal) dues rates and subscription prices except the 3.125 percent increase in dues adopted by the Executive Committee to be effective January 1, 1984. In this variant the surplus in current dollars in 1987 is $334 thousand, having declined from $370 thousand in 1984.

In the past, the main purpose of the net worth, when positive, has been to serve as a safety net. As a rule of thumb, a ratio of net worth to annual expenditures of 0.5 has been deemed adequate. In the main projection, the ratio rises steadily from 0.88 at the start of 1982 to 1.69 at the end of 1987.

Since 1976, the Executive Committee has pursued a policy of raising dues in nominal terms less than the rise in the general price level. Under this policy, dues rates after adjustment for inflation have fallen substantially. For the next several years, the Executive Committee expects to hold nominal dues constant. In real terms, dues will fall in proportion to the inflation rate.

Even under the dues policy described, financial resources will be more than sufficient for current activities. The Executive Committee therefore is considering whether there are any new activities the Association should undertake. Suggestions from members will be welcome.

At the present time, audited financial results for 1983 are not available. They will be published in the June 1984 issue of the *Review*. The accompanying table shows unaudited results for the first nine months of 1982 and 1983, audited results for the full year 1982, and the budgets for 1983 and 1984.

The Association is blessed with an extraordinarily capable accountant, Norma Ayres, whose outstanding work deserves acknowledgment.

RENDIGS FELS, *Treasurer*

Report of the Finance Committee*

The accompanying inventory summary lists the securities held by the American Economic Association as of December 31, 1983, with costs and market values as of that date. The total market value of the securities portfolio at year end was $3,225,960. After making adjustments for cash additions and withdrawals, we estimate that the Association's investment portfolio experienced a total investment return of +15.2 percent during 1983. This total portfolio return was the result of the bonds experiencing a +10.2 percent return and equities a +16.8 percent return. Looking at the broader sweep of the last twenty-four months, the Association's common stocks generated a +50.4 percent total return which compares with a +48.8 percent return for the Standard and Poor's 500 Index.

Reflective of the transition of the U.S. economy from recession to recovery and in anticipation of improving corporate profits, several portfolio changes were made last year. These included new commitments of individual issues held at year end in Federal National Mortgage Association, General Motors, John Harland, Shared Medical Systems, Molex, Northern Telecom, Overnite Transportation, Pneumo Corporation, and Rohm and Haas, and the elimination of individual holdings of Arco, Houston Industries, Standard Oil of Ohio, Warner Communications, Merck, Texas Utilities, Waste Management, Kodak, Baxter Travenol, MCI Communications, Pfizer, Phillip Morris, and Graphic Scanning. Fixed-income securities continued to have a medium-term orientation with an average weighted maturity of 7.4 years.

The Finance Committee at its December 1983 meeting retained its existing directive specifying a 50–75 percent operating range for the portfolio's equity ratio. The average maturity of fixed-income assets, other than those invested in the Stein Roe Bond Fund, will be no more than eight years. A maximum investment in the Bond Fund of up to 10 percent of the portfolio's assets was authorized.

RENDIGS FELS, *Chairman*

TABLE 1—INVENTORY SUMMARY AS OF DECEMBER 31, 1983

	Value	Percent	Estimated Income	Estimated Current Yield
Case Equivalents	110,566	3.4	9,729	8.8
Short-Term Securities	313,407	9.7	41,363	13.2
Medium-Term Securities	312,744	9.7	39,988	12.8
Long-Term Securities and Preferred Stocks	199,882	6.2	221,82	11.1
Convertible Securities	84,375	2.6	8,726	10.3
Equity Securities	2,204,986	68.4	35,618	1.6
TOTAL	3,225,960	100.0	157,606	4.9

*The Report of the Finance Committee is informational and is not an audited financial statement. Consequently, there may be some modest discrepancies between figures in the Report of the Finance Committee and the Auditors' Report which will appear in the forthcoming June 1984 issue of this *Review*.

TABLE 2—INVENTORY AND APPRAISAL AS OF DECEMBER 31, 1983

	Amount	Price	Value	Unit Cost	Total Income	Estimated Income
Cash Equivalents and Fixed-Income Securities (29.0 percent)						
Cash Equivalents (0–1 year)(11.8 percent)						
Cash			(824)		(824)	(73)
Steinroe Cash Reserves, Inc.	111,389	1	111,390	1	111,390[a]	9,802
Total Cash Equivalents			110,566		110,566	9,729
Other Short-Term Securities (1–5 years)(33.5 percent)						
Fed Home Loan Bks (13.900 07/25/85)	50,000	104	52,219	100	50,000	6,950
Fed Farm Cr Bks (15.800 01/20/86)	50,000	109	54,281	100	50,000	7,900
Fed Home Loan Bks (15.300 02/25/86)	50,000	108	53,875	100	50,000	7,650
Fed Natl Mtg Assn (14.625 06/10/86)	50,000	107	53,563	100	49,953	7,313
Fed Home Loan Bks (11.400 10/25/88)	50,000	99	49,531	100	49,953	5,700
Fed Natl Mtg Assn (11.700 11/10/88)	50,000	100	49,938	100	49,938	5,850
Total Other Short-Term Securities			313,407		299,844	41,363
Medium-Term Securities (5–10 years)(33.4 percent)						
Fed Home Loan Bks (15.100 02/27/89)	50,000	112	56,000	100	50,000	7,550
Crdthrift Finl Nt (10.250 04/15/89)	50,000	92	46,098	84	42,039	5125
US Treas Nts (14.500 07/15/89)	50,000	111	55,656	99	49,733	7,250
Florida Pwr 1st (13.300 11/01/90)	50,000	102	51,208	95	47,598	6,650
Duke Pwr 1st Mtg (15.125 03/01/91)	50,000	109	54,376	99	49,625	7,563
Fed Home Ln Bks (11.700 04/27/92)	50,000	99	49,406	100	50,000	5,850
Total Medium-Term Securities (10 years)			312,744		288,995	39,988
Long-Term Securities (More than 10 years)(21.3 percent)						
Hydro-Quebec Debentures (11.250 10/15/09)	50,000	99	49,593	96	48,031[a]	5,625
Steinroe Bond Fund	17,764	8	150,289	8	150,000[a]	16,557
Total Long-Term Securities			199,882		198,031	22,182
TOTAL CASH AND FIXED-INCOME SECURITIES			936,599		897,436	113,262
Convertible Securities (Limited Risk) (2.6 percent)						
Convertible Bonds (100.0 percent)						
Datapoint CV (8.875 06/01/06)	50,000	73	36,250	67	33,250	4,438
Barnett Banks CV (12.250 12/15/06)	35,000	138	48,125	100	35,000	4,288
TOTAL CONVERTIBLE SECURITIES (Limited Risk)			84,375		68,250	8,726
Equity Securities (68.4 percent)						
Energy Services (3.5 percent)						
Schlumberger Ltd	900	50	45,000	31	28,238	936
Food, Beverages, and Tobacco (3.5 percent)						
General Cinema	1,000	45	45,000	16	16,215	640
Publishing (9.9 percent)						
Amer Greetings Corp Cl A	2,000	27	53,000	13	26,125	840
Commerce Clearing House	600	60	36,000	49	29,100	1,032
Harland John H Co	1,000	37	37,000	38	38,450	760
			126,000		93,675	2,632
Major Chemicals (2.4 percent)						
Rohm & Haas	500	61	30,375	66	33,163	800
Drugs and Hospital Supplies (4.3 percent)						
Abbott Lab	1,200	45	54,300	12	14,292[a]	1,200
Machinery; Fabr Metal Product (4.3 percent)						
Diebold Inc	700	78	54,338	49	34,501[a]	840

(Continued)

TABLE 2—*(Continued)*

	Amount	Price	Value	Unit Cost	Total Income	Estimated Income
Computers and Office Equipment (8.8 percent)						
IBM	480	122	58,560	28	13,325[a]	1,824
Wang Labs Cl B	1,490	36	53,081	19	28,487	179
			111,641		41,812	2,003
Electrical Equipment (10.0 percent)						
Emerson Electric	700	67	46,550	61	43,036	1,610
General Electric	1,380	59	80,903	18	24,536[a]	2,760
			127,453		67,572	4,370
Electronics (6.8 percent)						
Molex	600	79	47,250	69	41,100	30
Northern Telecom Ltd	1000	39	39,000	40	40,240	400
			86,250		81,340	430
Automotive (2.9 percent)						
General Motors	500	74	37,188	64	31,795	2,000
Aerospace and Transport Equipment (3.2 percent)						
Pneumo Corp	1,400	29	40,425	28	38,626	700
Air Transport, Trucking, and Shipping (4.0 percent)						
Overnite Trans	1,700	30	51,425	26	44,047	952
Broadcasting and Communications (2.7 percent)						
Metromedia	1,000	35	35,000	14	14,274	760
Distribution (19.1 percent)						
Macy R H & Co Inc	900	52	47,025	43	38,421	720
McDonalds	900	71	63,450	40	36,315	900
Wal-Mart Stores	2,500	39	97,500	9	22,094	350
Dennys Cv (9.500 10/15/07)	30,000	120	36,000	100	30,000	2,850
			243,975		126,830	4,820
Banks; Savings and Loans (2.9 percent)						
Citicorp	1,000	37	37,125	27	26,595	1,880
Other Financial Services (2.9 percent)						
Federal Natl Mtge Assn	1,600	23	36,800	20	32,608	256
Services (8.9 percent)						
Humana Inc	2,666	26	69,983	9	24,962[a]	1,920
Shared Medical Systems	1,300	33	43,225	28	35,750	520
			113,208		60,712	2,440
Miscellaneous (28.8 percent)						
Steinroe Discovery Fund Inc	20,247	9	179,393	10	200,000[a]	
Steinroe Special Fund	8,248	18	146,253	13	107,695[a,b]	2,392
Steinroe Universe Fund Inc	32,746	18	603,837	12	404,122[a,b]	5,567
			929,483		711,817	7,959
TOTAL EQUITY SECURITIES			2,204,986		1,498,112	35,618
TOTAL SECURITIES AND CASH			**3,225,960**		**2,463,798**	**157,606**

Note: Total cost figure is low because basis on at least one security is missing.

[a] More than one cost basis.

[b] Represents investment cost; tax cost is higher.

Report of the Managing Editor
American Economic Review

I have no changes in editorial policy to report for the year 1983 as compared with 1982. We have been less than 100 percent successful in ensuring that authors have a final editorial decision within three months of the date that their manuscript is received at the editorial office, but the exceptional cases are mainly ones in which increased speed would be associated with less considered editorial assessments. On the whole, the operations of the editorial office have gone smoothly in all directions.

Operations

The recent history of manuscript submissions and papers published is shown in Tables 1 and 2. We received 932 papers in 1983. This is a more than 10 percent increase over 1982, and indicates that my earlier expectation of a leveling off in submissions was nonrational. The generally quick turnaround time at the *Review* seems to carry more weight with respective authors than does the manuscript submission fee.

The disposition of manuscripts received during 1982 and 1983 is shown in Table 3. The acceptance rate for 1983 is significantly lower than last year. I have no explanation for this, but I conjecture that it reflects a random variation either in the quality of papers submitted, or in editorial standards. No doubt the acceptance rate will rise again next year.

Our file of accepted papers as of December 31, 1983, contained 81 manuscripts; 29 of these will appear in March 1984, 30 or so in June 1984, and the remainder in September.

Additional information about editorial office processing lags is provided in Tables 4 and 5. The average inventory of manuscripts "in process" has again increased substantially as it has during each of the past two years. This appears to be a result not of

TABLE 1—MANUSCRIPTS SUBMITTED AND PUBLISHED, 1964–83

Year	Submitted	Published	Ratio of Published-to-Submitted
1964	431	67	.16
1965	420	59	.14
1966	451	62	.14
1967	534	94	.18
1968	637	93	.15
1969	758	121	.16
1970	879	120	.14
1971	813	115	.14
1972	714	143	.20
1973	758	111	.15
1974	723	125	.17
1975	742	112	.15
1976	695	117	.17
1977	690	114	.17
1978	649	108	.17
1979	719	119	.17
1980	641	127	.20
1981	784	115	.15
1982	820	120	.15
1983	932	129	.14

TABLE 2—SUMMARY OF CONTENTS, 1982 AND 1983

	1982			
	Number	Pages	Number	Pages
Articles	52	761	52	747
Shorter Papers, including Comments and Replies	68	407	77	383
Dissertations		21		24
Announcements and Notes Section		52		48
Index		10		10
Total		1251		1212

TABLE 3—DISPOSITION OF MANUSCRIPTS, 1982 AND 1983

	1982	1983
Manuscripts Received	820	932
Completed Processing:	694	732
Accepted	113	86
Rejected	581	646
Acceptance Rate	13.8%	9.2%
Currently in Process	126	200

TABLE 4—DISTRIBUTION OF EDITORIAL DECISION LAGS BETWEEN RECEIPT AND REJECTION, NOVEMBER 1, 1982–OCTOBER 31, 1983

Weeks to Rejection	Number of Manuscripts	Percent
0–4	349	.55
5–9	137	.22
10–14	93	.14
15–19	38	.05
20–24	18	.03
25–29	8	.01
30+	3	.00
Total	646	100.

TABLE 5—AVERAGE PUBLICATION LAGS, BY JOURNAL ISSUE

	Number of Weeks Lag		
Issue	Receipt to Acceptance	Acceptance to Publication	Receipt to Publication
March 1983	24	41	65
June 1983	20	34	54
September 1983	20	32	52
December 1983	22	39	59

slower reviewing, but rather of rising submission rates. There has been no significant change since last year in the distribution of editorial decision lags (Table 4).

The subject matter distribution of pages published in 1982 and 1983 is shown in Table 6. As indicated, a few changes have occurred, but none is in any way significant. Of course, the allocation to categories is somewhat arbitrary; what I would categorize as a contribution to industrial organization would be categorized by others as "welfare economics," "microeconomic theory," or "metaphysics" (the last category, to the regret of some, does not appear on our list).

TABLE 6—SUBJECT MATTER DISTRIBUTION OF PUBLISHED MANUSCRIPTS, 1982 AND 1983

	Published	
	1982	1983
General Economics and General Equilibrium Theory	6	12
Microeconomic Theory	17	15
Macroeconomic Theory	4	10
Welfare Theory and Social Choice	8	8
Economic History, History of Thought, Methodology	3	6
Economic Systems	2	4
Economic Growth, Development, Planning, Fluctuations	13	5
Economic Statistics and Quantitative Methods	6	5
Monetary and Financial Theory and Institutions	11	9
Fiscal Policy and Public Finance	8	8
International Economics	8	6
Administration, Business Finance	2	7
Industrial Organization	7	11
Agriculture, Natural Resources	3	4
Manpower, Labor Population	14	12
Welfare Programs, Consumer Economics, Urban and Regional Economics	8	7
Total	120	129

Expenses: Printing and Mailing

Table 7 shows the printing and mailing expenses for the four regular issues and for the *Papers and Proceedings* issue of the *Review* for 1983. As in earlier years, the *Papers and Proceedings* accounted for approximately 25 percent of total printing and mailing expenses. There have been only minor changes in costs during the preceding year, and I expect that situation to continue. Accordingly, expenses for 1984 are projected to be much the same as in 1983.

Papers and Proceedings

The sixth volume of the *Papers and Proceedings* to be prepared by the editorial staff of the *Review* appeared in May 1983. This year the task was handled by John Riley, the

TABLE 7—COPIES PRINTED, SIZE, AND COST OF PRINTING AND MAILING, 1983 *AER*

		Pages		Cost		
	Copies Printed	Net	Gross	Issue	Reprints	Total
March	28,000	256	302	$45,254.87	$1,822.89	$47,078
May	28,000	426	456	67,765.79	3,273.02	71,039
June	27,500	262	288	50,052.20	1,418.23	51,470
September	27,500	352	384	58,048.17	1,756.94	59,805
December[a]	27,500	342	392	55,826.86	2,000.00	57,827
Annual Misc.[b]						9,781
Total		1,638	1,824	$276,947.89	$10,271.08	$297,000

[a] Estimated.
[b] Estimated: based on costs of preparing mailing list, extra shipping charges, and storage costs of back issues.

associate editor, and by Wilma St. John and Theresa De Maria. I played no role in the process, nor do I intend to have anything to do with it during the coming year.

Last year I reported that, in an attempt to resolve some of the problems in adding the *Proceedings* to our regular work load, we would go directly from manuscript to page proofs in the 1983 *Proceedings*. That is what we did, and the experiment worked extremely well. Accordingly, we shall continue with that procedure in handling the 1984 volume.

Board of Editors

The board of Editors now consists of twelve members, chosen by the managing editor, with the approval of the Executive Committee of the Association. As in earlier years, the Board has been particularly helpful in dealing with comments on published articles and in reading papers that are the subject of complaint over the fairness or competence of referees. It goes without saying that I am grateful to members of the Board for their assistance and advice, for their generally supportive attitude, and for often useful criticism of particular actions of the managing editor or particular aspects of the operations of the *Review*.

Six members of the present Board will complete their terms of March 31, 1984: Gerald Bierwag, Thomas Cooley, Ronald Ehrenberg, Ted Frech, Laurence Kotlikoff, and Roy Weintraub. I thank them all for work well done, and I promise that I shall continue to call upon them in the future! I also want to thank the continuing members of the Board for services performed and in prospect: George Akerlof, Patricia Danzon, Jack Hirshleifer, Rick Mishkin, Sherwin Rosen, and Richard Schmalensee.

Acknowledgements

I wish to thank the associate editor, John Riley, and my office associates, Wilma St. John, Theresa De Maria, Lois Bagley, and Marcus Hennessy for dedicated performance that is frequently beyond the call of duty. My thanks also to our graduate mathematics consultant, Peter Swank. Finally, as in earlier years, I want to express my heartfelt thanks to the nearly 500 referees whose efforts make the publication of the *Review* possible.

J. M. Abowd
R. D. Adams
R. Aiyagari
G. Akerlof
A. Alchian
A. Alexander
B. T. Allen
W. R. Allen
J. E. Anderson
K. Arrow
O. Ashenfelter
B. K. Atrostic
R. Ayanian
C. Azariadis
R. E. Baldwin
M. N. Baily
P. Bardhan
D. P. Baron
A. P. Barten
Y. Barzel
L. Bassett
A. B. Batchelder
W. Baumol
J. R. Behrman
D. Bellante
B. L. Benson
T. C. Bergstrom
B. Bernanke
C. H. Berry
R. R. Betancourt
J. Bhagwati
K. B. Bhatia

G. O. Bierwag
J. F. O. Bilson
G. Bittlingmayer
F. Black
F. D. Blau
M. Blaug
M. I. Blejer
A. S. Blinder
N. E. Bockstael
P. Bohm
L. A. Boland
E. W. Bond
M. D. Bordo
G. Borjas
B. Boulier
N. Bowers
S. Bowles
K. Boyer
D. Bradford
S. D. Braithwait
W. H. Branson
M. Bray
G. F. Break
H. G. Brennan
T. F. Bresnahan
G. Brock
C. C. Brown
E. K. Browning
W. H. Buiter
C. Bull
M. Burstein
G. T. Burtless
G. C. Cain
G. Calvo
M. B. Canzoneri
J. A. Carlson
D. Carlton
J. Carr
F. Carrada
G. Chamberlain
S. Cheung
R. S. Chirinko
B. R. Chiswick
K. A. Chrystal
K. Clark
R. H. Coase
R. D. Coe
W. S. Comanor
T. F. Cooley
R. V. L. Cooper
B. Cornell
R. Cotterman
P. J. Coughlin
C. Cox
J. Cox
M. Crain
R. G. Crawford
V. Crawford
J. Cremer
H. D'Angelo
P. Danzon
M. Darby
R. Dardis
J. B. Davies
P. Davidson
L. De Alessi
A. V. Deardorff
W. D. Dechert
H. Demsetz
M. G. S. Denny
A. De Serpa
J. Devanso
A. De Vany
D. Dewey
P. A. Diamond
A. K. Dixit
P. B. Dixon
R. Dornbusch
G. K. Dow
A. Drazen
J. H. Dreze
D. Easley
F. H. Easterbrook
L. Ebrill
B. Eden
S. Edwards
R. G. Ehrenberg
B. C. Ellickson
D. T. Ellwood
W. Ethier
O. Evans
M. J. Ezrati
R. E. Falvey
E. F. Fama
G. Fane
H. S. Farber
D. J. Faurot
A. J. Fechter
G. Feder
J. Ferejohn
L. Fernandez
A. Field
S. Fischer
A. Fishlow
R. J. Flanagan
J. Frankel
W. J. Frazer
H. E. Frech III
K. R. French
J. A. Frenkel
B. Frey
D. Friedman
D. D. Friedman
M. Friedman
R. B. Friedman
R. Froyen
R. Frydman
V. R. Fuchs
D. Fudenberg
D. Fullerton
E. G. Furubotn
M. A. Fuss
J. Geweke
W. I. Gillespie
V. P. Goldberg
F. Gollop
R. J. Gordon
G. Gotz
E. Gramlich
P. Graves
S. I. Greenbaum
M. L. Greenhut
J. S. Greenlees
D. F. Greer
R. G. Gregory
J. M. Griffin
R. Gronau
E. Grossman
G. M. Grossman
H. I. Grossman
S. J. Grossman
J. L. Guasch
A. Guha
T. Gylfason
F. Hahn
C. Hall
R. Hall
J. Haltiwanger
K. Hamada
D. S. Hamermesh
P. Hammond
G. Hanoch
R. Hansen
E. A. Hanushek
D. Harrison
G. Harrison
M. Hashimoto
J. A. Hausman
F. Hayashi
R. Heiner
J. F. Helliwell
P. H. Hendershott
J. D. Hey
J. Hirshleifer
R. J. Hodrick
S. Hoenack
P. Hoffman
S. Hollander
C. A. Holt
I. Horowitz
P. Howitt
Y. M. Ioannides
D. Jaffee
M. Jensen
M. B. Johnson
T. R. Johnson
W. R. Johnson
C. Kahn
J. P. Kalt
M. Kamien
S. M. Kanbur
E. Karni
P. Kasliwal
M. L. Katz
K. Kimbrough
B. Klein
M. Kohn
R. J. Kopp
R. Kormendi
L. Kotlikoff
M. B. Krauss
M. Kreinin
P. J. Kuhn
H. Kunreuther
T. Kuran
F. E. Kydland
J.-J. Laffont
J. C. La Force
E. M. Landes
E. Lazear
E. Leamer
L.-F. Lee
K. Leffler
H. Leibenstein
A. Leibowitz
S. Leibowitz
J. P. Leigh
S. F. LeRoy

S. D. Lesnoy
M. Levi
D. Levine
W. A. Lewis
S. J. Leibowitz
D. M. Lilïen
L. A. Lillard
C. Lim
C. M. Lindsay
S. A. Lippman
R. E. Lucas, Jr.
S. Lustgarten
R. P. McAfee
J. J. McCall
B. T. McCallum
C. E. McClure
I. M. McDonald
J. McDonald
J. M. McDowell
R. I. McKinnon
A. McLaughlin
W. McManus
J. Machina
C. D. Macrae
T. Macurdy
S. P. Magee
J. H. Makin
B. G. Malkiel
N. G. Mankiw
H. G. Manne
J. Marchand
C. Marcuzzo
N. P. Marion
J. Markusen
T. Marschak
E. Maskin
W. E. Mason
R. W. Masulis
T. Mayer
W. Meckling
M. Melvin
R. C. Merton
R. A. Meyer
P. Mieszkowski
P. Milgrom
L. Mirman
F. Mishkin
H. Miyazaki
H. Mohring
R. L. Moore
C. Morris
D. Mortensen
L. Moses
J. Muellbauer
D. Mueller
P. Murrell
R. Musgrave
M. Mussa
D. M. G. Newbery
J. Newhouse
Y. K. Ng
V. D. Norman
D. C. North
W. E. Oates
H. Ohta
M. Olson, Jr.
A. E. Osborne
J. Ostroy
M. Ott
J. C. Panzar
D. O. Parsons
P. Pashigian
M. V. Pauly
D. K. Pearce
I. F. Pearce
S. Peltzman
J. H. Pencaval
J. M. Perloff
G. Perry
M. K. Perry
R. S. Pindyck
R. Piron
M. Plant
C. R. Plott
S. Polachek
A. M. Polinsky
R. A. Pollak
W. W. Pommerehne
C. Pope
J. M. Poterba
D. D. Purvis
J. M. Quigley
J. Quinn
R. Ram
S. Ranney
R. Rasche
E. J. Ray
A. Razin
M. Reder
J. D. Reid, Jr.
J. Reinganum
F. Rivera-Batiz
A. M. Rivlin
J. Roberts
M. Robinson
A. J. Robson
K. S. Rogoff
V. V. Roley
D. Roper
H. S. Rosen
K. T. Rosen
S. Rosen
M. R. Rosenzweig
J. J. Rotemberg
A. E. Roth
M. Rothschild
P. H. Rubin
R. Ruffin
J. Sachs
S. W. Salant
J. Salmon
S. C. Salop
P. A. Samuelson
W. F. Samuelson
T. Sandler
A. M. Santomero
T. J. Sargent
M. E. Satterthwaite
T. R. Saving
D. T. Scheffman
R. Schmalensee
F. M. Scherer
T. Schnelling
T. P. Schultz
R. M. Schwab
A. J. Schwartz
M. Schwartz
L. Seidman
C. Shapiro
P. Shapiro
L. Shapley
S. Shavell
H. M. Shefrin
W. Shepherd
R. J. Shiller
C. S. Shoup
J. J. Siegfried
J. Silvestre
P. Slovick
J. Smith
L. B. Smith
R. S. Smith
V. L. Smith
K. Sokoloff
L. C. Solomon
R. W. Solow
H. Somers
D. F. Spulber
D. O. Stahl II
R. Starr
M. Staten
R. Stevenson
G. J. Stigler
J. E. Stiglitz
E. Stokey
G. G. Storey
C. Stuart
A. H. Studenmund
L. H. Summers
R. Summers
V. Tanzi
P. Taubman
J. B. Taylor
L. Taylor
T. J. Teisberg
L. G. Telser
R. Thaler
A. Thomas
E. Thompson
J. Tirole
S. Titman
R. D. Tollison
R. Topel
R. M. Townsend
R. Tresch
J. Triplett
G. Tullock
S. Turnovsky
L. Tyson
D. Usher
R. A. Van Order
S. van Wijnbergen
H. Varian
K. Vaughn
R. Verrecchia
W. Vickrey
W. K. Viscusi
M. Waldman
N. Wallace
R. Wallace
R. N. Waud
M. Ward
R. L. Weil
H. M. Weingartner
B. Weingast
R. Weintraub
A. Weiss
L. Weiss

L. W. Weiss
B. A. Weisbrod
M. Weitzman
F. Welch
S. Wellisz
C. West
F. M. Westfield
J. Weymark
J. Whalley
D. J. White
J. Wilcox
L. L. Wilde
S. Wildman
T. D. Willett
O. E. Williamson
C. A. Wilson
G. C. Winston
S. G. Winter
A. D. Witte
G. Woglom
B. Wolfe
P. Wonnacott
S. A. Woodbury
S. Woodward
L. Young
B. Yu
R. Zeckhauser
T. Zorn

ROBERT W. CLOWER, *Managing Editor*

Report of the Managing Editor
Journal of Economic Literature

The general content and form of the *Journal* in 1983 remained unchanged from that established several years ago.

The documentation services of the *Journal* are carried on in Pittsburgh. The five departments of the *Journal* comprising these services are in charge of Naomi Perlman, the *Journal*'s senior associate editor. She has the help of Drucilla Ekwurzel, the assistant editor in the Pittsburgh office. Mrs. Perlman and the Pittsburgh staff also prepare the annual *Index of Economic Articles*. Volume XX of the *Index* covering 1978 was published during the current year. Mrs. Perlman's negotiations with Dialog Information Retrieval Service came to fruition and on-line computer access to the *JEL* and *Index* data bases became available in February 1983. The Pittsburgh office arranges for processing of computer tapes to Dialog. This contract also yields royalties to the Association. An additional report, prepared by Mrs. Perlman, covering the documentation services and other activities of the Pittsburgh office follows my report.

The Articles and Book Review departments of the *Journal* are edited in the Stanford office. Associate editor John Pencavel is in immediate charge of the Stanford staff and shares responsibility with the managing editor for the articles. Alexander Field serves as associate editor in charge of book reviews. Anne R. Saldich is the assistant editor in the Stanford office.

During 1983 (including the December issue soon to reach readers), the Journal published fourteen titles in the Articles Department. These included eight articles, one review article, and five substantial Notes. There were also two Communications. The books reviewed numbered 153. Details of the data published in the several documentation departments appear in Mrs. Perlman's following report.

I refer readers again to the statement of editorial objectives and policies set forth in an Editor's Note in the June 1981 issue, page 491. In accordance with these policies, the managing editor commissions the *Journal*'s expository, survey and review articles. But the *Journal* welcomes proposals for such articles. The managing editor also commissions book reviews.

On behalf of the Association, I should like to express my warm thanks to the 1983 Board of Editors, who helped plan and review the *Journal's* articles, and to the many other economists who served as referees. I acknowledge with special thanks the contributions of the following members who complete three-year terms on the Board this year: Alan S. Blinder, Donald J. Harris, John W. Kendrick, John Michael Montias, Roger G. Noll, Michael Rothschild, Isabel V. Sawhill, and Finis Welch.

Lyndis Rankin in Pittsburgh and Ann G. Vollmer and Anita Makler in Stanford continued on the *Journal* staff throughout 1983. The *Journal* is very grateful for their devoted and efficient work. Margaret Yanchosek, who joined the *Journal* in Pittsburgh in 1973, retired in August. Economists throughout the world, who depend on the accuracy and completeness of the *Journal's* documentation service, owe a great deal to her thorough and painstaking efforts.

MOSES ABRAMOVITZ, *Managing Editor*

Report of the Pittsburgh Office, *JEL*

The Pittsburgh office of the *Journal of Economic Literature* operates as an autonomous unit with complete responsibility for the following sections of the *Journal*: new books, an annotated listing; contents of current periodicals; subject index of articles in current periodicals; selected abstracts; and various indices including index of authors of articles in the subject index and index of authors of new books. In connection with the annotation section, we screen books for review for the Stanford office.

The Pittsburgh office also produces the *Index of Economic Articles* and the Economic Literature Index available on Dialog Information Retrieval Service.

The production process includes supervision of all stages of activities beginning with receipt of books and journals to the printed or on-line product (including arrangements for programming for the on-line file and for new indices). The printing prices, payments to Richard D. Irwin, Inc. for distribution, and royalty arrangements with Dialog are not handled by Pittsburgh, except for the paper price for the *Index of Economic Articles*.

I am pleased to report that in February 1983 the efforts of the Pittsburgh office finally came to fruition and the Economic Literature Index became available for on-line searches by arrangement with Dialog Information Retrieval Service. This Index includes all journal articles listed in the *Index of Economic Articles*, 1969–78, plus all articles listed in the *JEL*, 1979–83. We update the file with each quarterly issue of *JEL*. The file may be searched by subject classification (some 300 plus), author's name, title, individual words in the title, and geographic descriptor. Also in 1983, the 1978 *Index of Economic Articles* was published, containing 10,700 individual articles from 236 journals and 261 books.

A new index of authors of new books with annotations published in *JEL* 1983 was added. Hereafter, this index will be published quarterly.

There are a variety of problems and/or choices that have arisen in connection with the extension of journal coverage and with the extension of on-line access to annotations and abstracts. I believe it is necessary to have a small advisory committee or subcommittee of the Board of Editors with whom I can consult on these questions and on future information services.

The Association receives revenues from spinoff and extensions of the bibliographic activities of the Pittsburgh office, which I estimate at about $73,000 for 1983: *Index of Economic Articles*: $63,000; Economic Literature Index: $10,000. I anticipate that Dialog revenues will increase in 1984.

The Pittsburgh office continues to be well served by its long-term employees, Lyndis Rankin, consultant Asatoshi Maeshiro, and Drucilla Ekwurzel. Margaret Yanchosek, after ten years of similar devoted service, chose to take early retirement and has been replaced by Patricia Andrews. We also wish to thank our part-time employees Elizabeth Braunstein, Jang-Bong Choi, Joan Daley, Adriaan Dierx, A'Amer Farouqi, Aurelia Hooley, Nayyer Hussain, Barbara McGowan, and David M. McKibben.

NAOMI PERLMAN, *Associate Editor*

Report of the Director

Job Openings for Economists

For the third consecutive year, the number of new jobs listed declined from the previous year. Last year (1982), 1,659 new vacancies were advertised; this year only 1,489 new jobs were listed—a decline of 10 percent from 1982. Both academic and nonacademic listings decreased although the percentage decline in nonacademic jobs was greater. Table 1 shows total listings (employers), total jobs, new listings, and new jobs, by type (academic or nonacademic) for each issue of *Job Openings for Economists* in 1983.

Universities with graduate programs and four-year colleges continue to be the major sources of job listings. Together they constitute about 83 percent of total employers. Table 2 shows the number of employers by type for each 1983 issue.

TABLE 1—JOB LISTINGS FOR 1983

Issue	Total Listings	Total Jobs	New Listings	New Jobs
Academic				
February	78	149	59	111
April	51	80	44	60
June	24	48	23	47
August	35	89	34	84
October	146	358	129	311
November	137	313	137	313
December	200	457	87	187
Subtotal	671	1,494	410	1,113
Nonacademic				
February	16	54	12	35
April	15	56	13	46
June	13	49	11	39
August	16	69	14	59
October	24	105	18	74
November	18	51	18	51
December	36	131	20	72
Subtotal	138	515	106	376
TOTAL	809	2,009	516	1,489

TABLE 2—NUMBER AND TYPES OF EMPLOYERS LISTING POSITIONS IN *JOE* DURING 1983

Issue	Four-Year Colleges	Universities with Graduate Programs	Federal Government	State/Local Government	Banking or Finance	Business or Industry	Consulting or Research	Other	Total
February	43	35	4	1	2	–	7	2	94
April	22	29	2	4	3	1	3	2	66
June	13	11	2	3	1	2	3	2	37
August	7	28	1	3	5	3	2	2	51
October	46	100	9	1	3	2	7	2	170
November	45	92	2	2	6	1	7	1	155
December	86	113	11	4	8	1	10	3	236
TOTAL	262	408	31	18	28	10	39	13	809

TABLE 3—FIELDS OF SPECIALIZATION CITED: 1983

Fields[a]	February	April	June	August	October	November	December	Totals
General Economic Theory (000)	68	46	21	36	189	133	222	715
Growth and Development (100)	25	17	11	17	52	24	49	195
Econometrics and Statistics (200)	35	17	12	19	57	41	84	265
Monetary and Fiscal (300)	27	25	13	19	79	77	111	351
International Economics (400)	25	20	8	14	37	32	51	187
Business Administration, Finance, Marketing and Accounting (500)	44	23	9	14	43	30	64	227
Industrial Organization (600)	13	21	9	11	50	34	63	201
Agriculture and Natural Resources (700)	8	10	13	10	21	17	28	107
Labor (800)	10	12	5	5	31	31	43	137
Welfare and Urban (900)	9	5	7	12	30	29	54	146
Related Disciplines (A00)	6	1	–	2	6	1	3	19
Administrative Positions (B00)	2	6	4	7	14	13	27	73
TOTAL	272	203	112	166	609	472	799	2,623

[a] Fields of specialization codes are from the *Journal of Economic Literature.*

The field of specialization most in demand continues to be general economic theory. Generalists with a strong background in mathematics and statistics appear to be the type of economist that employers are seeking. The applied area of specialization seems to be of secondary importance. Table 3 shows the number of citations by field of specialization. General economic theory (000) led, followed by monetary and fiscal (300) and econometrics and statistics (200). This pattern has prevailed for the past several years.

Violet Sikes is almost solely responsible for the publication and distribution of *JOE.* I wish to express my great gratitude for the excellent job she continues to do.

C. ELTON HINSHAW, *Director*

Report of the Representative to the National Bureau of Economic Research

Research carried out in 1983 at the National Bureau of Economic Research continued to be organized in eight main programs: Economic Fluctuations (Robert Hall), Financial Markets and Monetary Economics (Benjamin Friedman), International Studies (William Branson), Labor Studies (Richard Freeman), Taxation (David Bradford), Development of the American Economy (Robert Fogel), Health Economics (Victor Fuchs and Michael Grossman), and Productivity and Technical Change (Zvi Griliches).

Work continued during 1983 on several large-scale projects, which bring together researchers from several of these programs. John Shoven is overall director of the Bureau's pensions project. A major phase of the Bureau's work (directed by David Wise) on the effects of pensions on labor market and retirement decisions concluded during 1983, but further research on a wide range of pension-related issues is continuing.

Another major project, centering on the role of Government Budget and the Private Economy, consists of major efforts in the following areas: the impact of taxation on such behavior as charitable contributions (directed by Charles Clotfelter); measuring and analyzing the growth of government spending at the state and local level; the analysis of state and local government's role in the economy (directed by Harvey Rosen); an analysis of the impact of transfer programs, studies of the compensation of public sector employees (directed by David Wise), and the impact of public sector unionization (directed by Richard Freeman); and an analysis of government debt and deficits and their impact on the private sector (directed by David Bradford and Benjamin Friedman).

The third project, Productivity and Industrial Change in the World Economy, likewise has several major parts. William Branson and David Richardson are directing a project on international economic policy. Research on trade relations and trade policy is directed by Robert Baldwin. Richard Marston leads a group studying international macroeconomic coordination. Colin Bradford is directing a study of trade relations with Asian countries. Jacob Frenkel is directing research on exchange rate changes.

Completed during 1983 was the project under the direction of Richard Freeman on black inner city youth unemployment. Also concluded was the phase of Benjamin Friedman's research project on the changing roles of debt and equity finance dealing with corporate capital structures in the United States. Zvi Griliches and M. Ishaq Nadiri are continuing their study of research and development. Research, directed by Robert Gordon, on the changing nature of the business cycle will culminate in a conference in 1984.

Bureau conferences (and organizers) in the United States and abroad in 1983 included: "General Equilibrium" (John Shoven); "Corporate Capital Structures in the U.S." (Benjamin Friedman); "Pensions, Labor, and Individual Choice" (David Wise); "Trade Policy" (Robert Baldwin and J. D. Richardson); "Economics of Trade Unions" (Daniel Hamermesh); "International Seminar on Macroeconomics" (Georges de Menil and Robert Gordon); "Macroeconomics" (Robert Hall); "Recent Issues and Initiatives in U.S. Trade Policy" (Robert Baldwin); "Trade Policy" (William Branson and Alvin Klevorick); "Aspects of International Capital Mobility" (Maurice Obstfeld); "Inner City Black Youth Unemployment" (Richard Freeman); "International Capital Mobility and the Coordination of Monetary Rules" (Richard Marston); "Productivity in Health" (Victor R. Fuchs); "Fifth Annual Research Conference"; "Economics of the U.S. Retirement Income System" (Zvi Bodie, John Shoven, and David Wise); "Income and Wealth: Horizontal Equity Uncertainty and Measures of Well-Being" (Martin David and Timothy Smeeding).

In 1983, the following NBER books were published by the University of Chicago Press: *Behavioral Simulation Methods in Tax Policy*

Analysis (Martin Feldstein, ed.); *Trade and Employment in Developing Countries, 3: Synthesis and Conclusions* (Anne O. Krueger, ed.); *The International Transmission of Inflation* (Michael Darby, James Lothian, Arthur E. Gandolfi, Anna Schwartz, and Alan C. Stockman); *Financial Policies of the World Capital Market: The Problem of Latin American Countries* (Rudiger Dornbusch, Pedro Aspe Armella, and Maurice Obstfeld, eds.); *The Measurement of Labor Cost* (Jack E. Triplett, ed.); *Exchange Rates and International Macroeconomics* (Jacob A. Frenkel, ed.); *Financial Aspects of the United States Pension System* (Zvi Bodie and John Shoven, eds.); *Pensions in the American Economy* (Laurence J. Kotlikoff and Daniel E. Smith, eds.).

Additional NBER books which will be published by the University of Chicago Press in 1984 include the following: *R&D, Patents, and Productivity* (Zvi Griliches, ed.); *International Tax Comparisons* (Mervyn A. King and Don Fullerton, eds.); *Exchange Rate Theory and Practice* (John F. O. Bilson and Richard C. Marston, eds.); *Economic Transfers in the United States* (Marilyn Moon, ed.); *Retrospective on the Classical Gold Standard, 1821–1931* (Michael D. Bordo and Anna Schwartz, eds.); *The Structure and Evolution of Recent U.S. Trade Policy* (Robert E. Baldwin and Anne O. Krueger, eds.).

Over 230 participants, representing seventy-nine universities and other organizations in the United States and abroad, met in Cambridge in July and August for the Bureau's sixth annual Summer Institute. Six NBER programs held workshops and seminars: Economic Fluctuations, Financial Markets and Monetary Economics, International Studies, Labor Studies, Productivity, and Taxation.

The Business Cycle Dating Group identified November 1982 as the trough which signified the end of the recession.

The President of the Bureau is Eli Shapiro of M.I.T. The Chairman of the Board is Walter Heller of the University of Minnesota. Further information about the Bureau and its activities may be obtained from its publications, the *Digest* and the *Reporter*, or from David G. Hartman, Executive Director, 1050 Massachusetts Avenue, Cambridge, MA 02138, or from the undersigned at Johns Hopkins University. I am happy to acknowledge the assistance of Dr. Hartman in the preparation of this report.

CARL F. CHRIST, *Representative*

Report of the Representative to the U.S. National Commission for UNESCO

The attention of the U.S. National Commission for UNESCO was dominated by continued discussion of the Commission's future in the face of a threatened cutoff of State Department funds to support operation of the Commission, and later in the year by increasing indications that U.S. government dissatisfaction with UNESCO itself was intensifying. These issues crowded out the usual amount of consideration of substantive programs.

At the February 1983 meeting, the main reason for terminating support for the National Commission, according to Assistant Secretary of State for International Organization Affairs Gregory Newell, was that all government agencies were undergoing personnel reduction and expenditure cuts, and the State Department had to share in these cuts. He explained that he would arrange for a support staff in his Bureau to work part or full time on UNESCO matters and arrange for another bureau in the State Department to continue the U.S. contribution to UNESCO's "Man and the Biosphere" program. As Chairman James Holderman put it, the Commission's options were thus reduced to a range between trying to generate private financial resources and "dying with some sense of grace." A number of commissioners expressed objections to this proposed decision.

An Ad Hoc Committee, set up by the Commission in 1982 to determine how to continue the assessment of U.S. participation in UNESCO, reported among other things that commissioners representing nongovernmental organizations who had expressed their views on U.S. participation in UNESCO's work and on what the Commission's role should be, thought the Commission should oppose more actively the efforts of UNESCO members to enact rules guiding the operation of world news media, encourage the United States to play a more active role and not merely to limit damage on this and other issues of concern to the United States, coordinate views and comments of American professional communities represented on the Commission, review and formulate advice to the State Department on the position that the U.S. government should take with respect to UNESCO program activities, and advise on other aspects of U.S. participation, such as representation on U.S. delegations to UNESCO meetings and conferences.

Following a general discussion, the Commission requested its Executive Committee to study alternative arrangements for operating the Commission, including its financing, with a view to strengthening the links with and support from private organizations, and to submit recommendations for action in time for the Commission's next annual meeting.

The Commission heard reports on the highlights of the Fourth Extraordinary Session of the UNESCO General Conference, held in Paris in October 1982, from the U.S. Ambassador to UNESCO, Jean Gerard, and the Commission's Vice Chairman, John Fobes. Ambassador Gerard presented and commented on UNESCO's second medium-term plan covering the years 1984–89, which she thought both too ambitious and containing "ideological pitfalls." For example, it directed the Director-General to analyze global problems not only in specific fields of UNESCO competence, but also to cover "the socio-political and historical aspects" of these problems. She noted that traditional human rights appeared to be subordinated to collective "people rights," and that there were references to the desirability and feasibility of "adapting human rights to certain socio-economic contexts and to the specific needs of certain social categories." The chapter on development placed heavy blame for the problems of poor countries on the world market economy and transnational corporations.

The meeting of the Commission broke up into several sectoral groups, all of which reported back to the Commission. The Social

Sciences group's view was that (1) it is important to strengthen the participation of the U.S. social science community in UNESCO's social science program because the United States needs to keep abreast of new social science research in developing countries, and UNESCO provides an important vehicle for doing so; (2) now that the International Social Science Council has new leadership, it is important to establish better connections with it; (3) the Social Sciences Committee should continue to improve its linkage with the Social Science Research Council in New York and other social science groups in the United States. This calls for transmission of information about international social science developments, which requires adequate staffing of the Commission's secretariat.

Commissioner Nancie Gonzalez of the American Anthropological Association was chosen to chair the Social Sciences Committee to replace retiring Commissioner Lawrence Finkelstein.

In 1983 a symposium organized jointly by the U.S. and Canadian National Commissions for UNESCO met in Quebec to assess the fundamental problems, strengths, and challenges of the social sciences in North America. The proceedings of this symposium are about to be published by the University of Ottawa.

The Commission assisted in preparing the U.S. delegation to a UNESCO intergovernmental Conference on Education for International Understanding, Cooperation, and Peace and Education Relating to Human Rights, held in Paris in April 1983.

Under the auspices of the Commission an International Symposium on Communications and World Development was held at the University of South Carolina in October 1983.

Among studies published by UNESCO in 1983 were *Quality of Life: Problems of Assessment and Measurement*, the fifth in a series of socioeconomic studies and described as three studies that illustrate the different kinds of considerations encountered in research on the quality of life, and *Cost and Effectiveness Overview and Synthesis*, the third volume in *The Economics of New Educational Media.*

The Commission expected that the main item on the agenda of its forty-seventh meeting, held in December 1983, would be the report of its Ad Hoc Committee to consider the organizational structure and general functioning of the Commission itself. When the meeting actually convened, this report was pushed into second place by the question of whether the United States would continue its membership in UNESCO. As I reported a year ago, the Commission had held a special meeting in 1982 to make its own critical assessment of U.S. participation in UNESCO. This activity reflected the sense of tension and concern about the American relationship with UNESCO. This concern had apparently intensified, at least on the part of the U.S. government. In August 1983 the government launched its own review of U.S. participation in UNESCO, including an examination of major UNESCO programs, the agency's effectiveness as a vehicle for promoting U.S. interests, international cooperation and development, and including also an examination of U.S. options regarding the relationship to UNESCO, including the possibility of withdrawing from it.

Assistant Secretary Newell, addressing the Commission at this meeting, listed the specific concerns that have led the administration to reconsider U.S. membership. He mentioned a number of systemic problems of U.N. agencies including anti-Western bias; intense hostility to the State of Israel, reflected in numerous attempts to exclude it from U.N. agencies and their meetings; "statism," or the general tendency to have problems resolved by the state (e.g., the "New World Communications Order"); and highly inflated budgets (to which the United States contributes 25 percent) and generally inefficient management. The second and fourth of these were common to all U.N. agencies, he said, but UNESCO was even more political and less efficient than most of the others and its budgets were less restrained. The Commission was also addressed by Lawrence Eagleburger, Undersecretary of State for Political Affairs, who stressed that the decision was under intensive review but had not yet been made. He said that, whatever the decision, the present situation would not be continued; if the United States stays in UNESCO, it will intensify its activities. If we leave, of course things will be different, but we will continue to be concerned with cultural affairs. The government recognizes the danger that our withdrawal from UNESCO would leave the Soviet Union with a virtual monopoly of influence in UNESCO.

In the ensuing discussion, a commissioner who had been involved in both the U.S. withdrawal from and return to the International Labor Organization questioned that the withdrawal accomplished anything; it permitted several undesirable actions to be taken during our absence, and the return of the United States to the ILO was complicated.

The Chairman told the Commission that he had transmitted to Assistant Secretary Newell the views of the member organizations who had expressed ideas about continuing U.S. membership and that Secretary Newell had asked that some negative views be put in to "give the report credibility."

A resolution favoring continuation of U.S. membership was proposed and, after some discussion, was adopted by a vote of 41 to 8.

It was reported that the opening days of the twenty-second UNESCO General Conference, held in Paris October 25–November 26, 1983, were dominated by condemnation of the United States for its intervention in Grenada, another example (along with hostility to Israel) of the politicization of UNESCO.

The Commission next considered the report of the committee it had appointed earlier to consider and make recommendations about the future of the Commission. This report, which had been adopted by the Executive Committee, proposed reducing the number of commissioners from a maximum of 100 with a chairman chosen by the Commission to a maximum of 40 with a chairman appointed by the Secretary of State. The proposed commission would have a maximum of 15 commissioners at large appointed by the Secretary of State and a maximum of 25 institutional and individual members selected from among their own ranks, with no more than five from each of the five sectoral areas (i.e., education, the natural and social sciences, culture and the humanities, communications, and "other" socioeconomic aspects of human concerns and the status of women). In addition to commissioners, the new body would include an unlimited number of members, most representing institutions and organizations, and a second membership class consisting of individual members, reserved for individuals who have no formal affiliation with nongovernmental organizations but who wish to be part of the Commission's deliberations and work. Institutional members would be charged a membership fee of $100 and individual members a fee of $50 a year. Members attending the proposed annual meeting of the new commission would be permitted to participate in discussion but not to vote.

As your representative, I should report that I expressed opposition on principle to the proposal that any individual or organization, regardless of qualifications, should be able, by payment of a fee,. to become a member of a body intended to advise the U.S. government.

Discussion of these and other aspects of the report on the future of the Commission revealed enough evidence of desire for reconsideration and perhaps modifications to make clear that the report could not be acted upon during the meeting.. The chairman invited commissioners to send their comments and suggestions for changes in writing to him for transmission to the Executive Committee.

In the last week of 1983, the U.S. government announced that it had given the Secretary-General of UNESCO the required one-year's notice that it would withdraw from the organization on January 1, 1985, but that it had retained its right to rejoin, and regarded its departure as temporary.

WALTER S. SALANT, *Representative*

Report of the Representatives to the Council of Professional Associations on Federal Statistics

The Council of Professional Associations on Federal Statistics (COPAFS) was established late in 1980 as a means to provide timely, systematic information on developments in federal statistics to the professional associations, to stimulate discussion and response by the members of those associations to emerging issues, and to serve as a mechanism for bringing the views of the professions to bear on decisions that would ultimately affect federal statistical products.

Members of the American Economic Association were instrumental in the formation of COPAFS and the AEA was one of the original founding members. Other associations that now comprise the Council include the American Agricultural Economics Association, American Association for Public Opinion Research, American Political Science Association, American Public Health Association, American Sociological Association, American Statistical Association, Association of Public Data Users, Federal Statistics Users' Conference, National Association of Business Economists, Population Association of America, and the Society of Actuaries.

Issues addressed by the Council during the first three years of its operation encompass a broad array of concerns: budgets for federal programs which were in jeopardy as a consequence of fiscal constraints and initiatives to reduce the burden of reporting, resources for the governments's statistical policy and coordination activities, organizational changes affecting the status and integrity of statistical agencies, access to federal statistical products, and the substance of ongoing and proposed statistical programs.

The COPAFS representatives have met quarterly to receive briefings on major developments affecting the quality, integrity, and availability of federal statistical products, and to discuss opportunities for action by the member associations. Administration and congressional officials concerned with statistical policy as well as heads of virtually all U.S. government statistical agencies or programs (agriculture, BLS, census, education, health, justice, and taxation) have made presentations at those meetings and exchanged news with COPAFS representatives.

In addition, the Council has instituted a monthly newsletter designed to keep member associations apprised of important developments between meetings. The newsletter also has become a major vehicle for communicating with representatives of the administration and the Congress, as well as with members of the press and a variety of other concerned individuals and organizations.

The Executive Director of COPAFS, Katherine Wallman, as well as representatives of member associations, have spoken by invitation at congressional hearings, annual meetings of a number of the member groups, and at various other forums. COPAFs also has provided information about statistical program concerns of users and effects of cutbacks and organizational shifts to members of the press, public interest groups, trade and nonmember professional associations, and other interested parties. This in part has stimulated increased coverage in major publications such as the *New York Times*, *Wall Street Journal*, *Washington Post*, *Boston Globe*, *Business Week*, *Science*, and others, as well as in a number of professional and trade association publications.

Although not widely known within the AEA, COPAFS activities have benefitted empirical analyses and research of AEA members by protecting and enhancing the integrity, quality, and scope of U.S. government statistics. The AEA representatives to COPAFS urge that the AEA reaffirm its support for the work of the Council, and that budgetary funding be provided comparable to that furnished by other major organizations.

GARY FROMM; JOHN H. CUMBERLAND,
Representatives

The Committee on the Status of Women in the Economics Profession

Women are a growing presence in economics classes and in the economics profession. Among undergraduate economics majors and in undergraduate economics courses, 30 percent of the students are now women, as compared with 15 percent 10 years ago, in 1973. Women are now 21 percent of the graduate students pursuing the Ph.D. degree, as compared with 12 percent ten years earlier. Some progress is also being made in faculty representation for women economists. However, it is still the case that the higher one looks in the professional hierarchy, the fewer women one finds. In academe, where we have information in some detail, the situation can be summarized:

Women as a Percentage of:	1973	1983
All undergraduates	44	52
In Economics:		
Undergraduate majors	15	30
Ph.D. students	12	21
Ph.D. degrees awarded	8	14
Assistant Professors	9	16
Associate Professors	6	11
Full Professors	3	4

Some of the current disparity in the extent of women's representation in the bottom as opposed to the top of the hierarchy is caused

TABLE 1—DISTRIBUTION OF FULL-TIME FACULTY, BY TYPE OF INSTITUTION, ACADEMIC YEAR 1982–83

	Chair's Group			Other Ph.D.			Only M.A. Departments			Only B.A. Departments		
		Female			Female			Female			Female	
	Total	No.	Percent	Total	No.	Percent	Total	No.	Percent	Total	No.	Percent
Existing												
Professor	634	14	2.2	921	25	2.7	223	19	8.5	349	26	7.4
Associate	256	18	7.0	470	30	6.4	319	89	27.9	313	19	6.1
Assistant	343	44	12.8	512	68	13.3	215	56	26.0	401	66	16.5
Instructor	52	11	21.2	80	19	23.8	111	15	13.5	119	25	21.0
Other	40	7	17.5	50	7	14.0	117	94	80.3	38	4	10.5
New Hires												
Professor	5	0	0	6	0	0	3	0	0	8	0	0
Associate	11	1	9.1	5	2	40.0	4	0	0	14	0	0
Assistant	58	7	12.1	90	13	14.4	36	5	13.9	76	16	21.1
Instructor	16	2	12.5	33	5	15.2	9	3	33.3	19	11	57.9
Other	4	1	25.0	7	1	14.3	6	3	50.0	13	0	0
Promoted To Rank (1981–82)												
Professor	21	1	4.8	31	2	6.5	16	1	6.3	19	2	10.5
Associate	31	4	12.9	45	7	15.6	18	4	22.2	31	3	9.7
Assistant	3	0	0	9	0	0	2	1	50.0	21	4	19.0
Tenured at Rank (1981–82)												
Professor	2	0	0	4	0	0	2	0	0	41	35	85.4
Associate	22	3	13.6	32	3	9.4	12	3	25.0	38	12	31.6
Assistant	2	1	50.0	3	2	66.7	4	1	25.0	17	1	5.9
Other	2	0	0	2	0	0	0	0	0	1	0	0
Not Rehired												
Professor	27	0	0	35	1	2.9	9	2	22.2	9	1	11.1
Associate	10	1	10.0	17	1	5.9	6	1	16.7	6	1	16.7
Assistant	40	2	5.0	55	5	9.1	27	8	29.6	46	6	13.0
Instructor	10	3	30.0	21	5	23.8	2	1	50.0	21	4	19.0
Other	6	1	16.7	6	1	16.7	0	0	0	9	3	33.3

TABLE 2—PREVIOUS ACTIVITY OF NEW HIRES AND CURRENT ACTIVITY OF THOSE NOT REHIRED BY TYPE OF INSTITUTION AND SEX, ACADEMIC YEAR, 1982–83

	Previous Activity of New Hires				Current Activity of Not Rehired			
	Male		Female		Male		Female	
	No.	Percent	No.	Percent	No.	Percent	No.	Percent
Chair's Group	88	100.0	18	100.0	76	100.0	3	100.0
Faculty	20	22.7	4	22.2	36	47.4	3	100.0
Student	59	67.1	12	66.7	2	2.6	0	0
Government	3	3.4	1	5.6	12	15.8	0	0
Bus., Banking, Research	4	4.6	0	0	15	19.7	0	0
Other	2	2.3	1	5.6	11	14.5	0	0
Other Ph.D.	137	100.0	31	100.0	108	100.0	9	100.0
Faculty	40	29.2	7	22.6	57	52.8	7	77.8
Student	83	60.6	20	64.5	6	5.6	1	11.1
Government	7	5.1	2	6.5	13	12.0	1	11.1
Bus., Banking, Research	4	2.9	1	3.2	15	13.8	0	0
Other	3	2.2	1	3.2	17	15.7	0	0
M.A. Departments	52	100.0	13	100.0	41	100.0	8	100.0
Faculty	20	38.5	3	23.1	23	56.1	0	0
Student	20	38.5	9	69.2	3	7.3	1	12.5
Government	3	5.8	0	0	0	0	0	0
Bus., Banking, Research	3	5.8	0	0	8	19.5	1	12.5
Other	6	11.5	1	7.7	7	17.1	6	75.0
B.A. Departments	158	100.0	44	100.0	77	100.0	14	100.0
Faculty	56	35.4	8	18.2	35	45.5	3	21.4
Student	74	46.8	27	61.4	6	7.8	2	14.3
Government	5	3.2	0	0	5	6.5	0	0
Bus., Banking, Research	18	11.4	7	15.9	10	13.0	2	14.3
Other	5	3.2	2	4.6	21	27.3	7	50.0

by inevitable lags, as the increased number of women economists starting their professional lives move through their professional life cycle. However, we would be naive if we were to believe that this disparity will cure itself in time without special effort. We have the unhappy example of some of the other professions, where, unlike economics, women have always been well represented at the bottom and where they continue to have poor representation at the top.

The importance of increasing the pitifully small number of women economists in the top ranks of the profession is well expressed in the following comment by Cynthia Fuchs Epstein, the sociologist who has been the closest student of the place of women in the professions:

> Until some reasonable ratio is developed, the tiny number of women who have been successful are destined to be regarded as pathological and gender anomalies. In addition, because women are not generally counted among the successful, all women are regarded as deficient. Thus, women outside as well as inside the professions and occupations are regarded as second-class citizens, as incompetents dependent on males to make the important decisions; as giggling magpies who will contaminate the decorum of the male luncheon clubs and bars; as persons who can't be trusted to be colleagues.

One event taking place in 1983 was the completion of Alice Rivlin's term of service as Director of the Congressional Budget Office. Rivlin took over as Director on the first day of the CBO's existence, and built it up from scratch into a respected source of competent, timely and unbiased analysis and information for the Congress and, indeed, for all those interested in government policymaking. In a profession under fire, she was

TABLE 3—DISTRIBUTION OF SALARY FOR WOMEN FACULTY BY TYPE OF DEPARTMENT AND TIME IN RANK, ACADEMIC YEAR, 1982–83

Relative Salery for Rank	All Women		Time in Rank			
	Number	Percent	Total Percent	Above Median	At Median	Below Median
All Departments	406	100.0	100.0	30.5	42.1	27.3
Salary above Median	137	33.7	100.0	52.6	27.7	19.7
Salary at Median	131	32.3	100.0	14.5	73.3	12.2
Salary below Median	138	34.0	100.0	23.9	26.8	49.3
Ph.D., Chair's Group	75	100.0	100.0	33.3	45.3	21.3
Salary above Median	23	30.7	100.0	43.5	30.4	26.1
Salary at Median	25	33.3	100.0	16.0	68.0	16.0
Salary below Median	27	36.0	100.0	40.7	37.0	22.2
Ph.D., Other	143	100.0	100.0	35.7	36.4	28.0
Salary above Median	51	35.7	100.0	54.9	25.5	19.6
Salary at Median	38	26.6	100.0	18.4	71.1	10.5
Salary below Median	54	37.8	100.0	29.6	22.2	48.1
M.A. Departments	60	100.0	100.0	33.3	41.7	25.0
Salary above Median	23	38.3	100.0	65.2	30.4	4.3
Salary at Median	19	31.7	100.0	21.1	63.2	15.8
Salary below Median	18	30	100.0	5.6	33.3	61.1
B.A. Departments	128	100.0	100.0	21.9	46.9	31.3
Salary above Median	40	31.3	100.0	47.5	27.5	25.0
Salary at Median	49	38.3	100.0	8.2	81.6	10.2
Salary below Median	39	30.5	100.0	12.8	23.1	64.1

virtually unique in the respect accorded her work. Rivlin and the staff she organized and directed were able unerringly to thread the political minefields of Capital Hill without compromise to their professional performance on the technical level.

While CSWEP is proud of Rivlin's performance as an economist, we also wish to call attention to her exemplary performance as an employer of economists. Out of a CBO professional staff of 166, women currently hold 58 professional jobs, or 35 percent.

Rivlin will be the Director of Economic Studies at The Brookings Institution, where she will have ample scope to improve the representation of women economists.

We commend to Rudolph G. Penner, Rivlin's successor as Director at CBO, the keeping of the now-established CBO tradition of open opportunities for women economists. We are pleased to report that among his initial acts has been the promotion of Rosemary Marcuss to be Assistant Director for Tax Analysis. At the Assistant Director level, Marcuss joins Nancy M. Gordon, who is Assistant Director for Human Resources and Community Development.

In contrast to CBO's hospitality to the talents of women economists was the action of Martin Feldstein, who in a well-publicized move, brought an all-male professional group with him to the Council of Economic Advisers. In both Democratic and Republican administrations in the past, the Council has employed a number of women economists as Council Members and on the senior staff. Feldstein's response to CSWEP's remonstrance was that he brought people he knew could do the job, and that if CSWEP could tell him of some women who could do the job he would be glad to consider them. We understand that CSWEP's protest has resulted in the subsequent hiring of a woman with a BA in economics onto the junior CEA staff.

Back at Harvard, where he was a professor, and the National Bureau of Economic Research, of which he was president, Feldstein left behind him two organizations in which women economists with senior roles are unusually rare, a fact possibly contributing to his lack of knowledge of women economists who can do the job. CSWEP is concerned about this rarity, and is consider-

TABLE 4—DEGREES GRANTED IN ECONOMICS BY TYPE OF DEPARTMENT AND SEX, ACADEMIC YEAR 1982–83

	All	Ph.D. Departments			M.A.	B.A.
Number of:	Depts.	Total	Chair's	Other	Depts.	Depts.
Departments	377	120	44	76	45	212
Ph.D.s	867	867	378	489	–	–
Female	122	122	50	72	–	–
Percent Female	14.1	14.1	13.2	14.7	–	–
M.A.s	1,705	1,529	538	991	176	–
Female	403	368	122	246	35	–
Percent Female	23.6	24.1	22.7	24.8	19.9	–
B.A.s	18,712	12,579	5,206	7,373	1,124	5,009
Female	5,687	3,681	1,535	2,146	346	1660
Percent Female	30.4	29.3	29.5	29.1	30.8	33.1
Other	287	280	39	241	2	5
Female	82	79	9	70	1	2
Percent Female	28.6	28.2	23.1	29.0	50.0	40.0

Note: Some departments do not report students by sex, and the figures in the table contain some allocations. The percentages, however, were not affected.

ing ways in which Harvard and NBER can be encouraged and assisted to allow more women economists into their valuable colleagueship.

CSWEP is also concerned about women economists' access to publication in professional journals and to participation in the programs of professional meetings. Research has shown that professional articles do better in the refereeing process if they are signed with a male name. We therefore believe that the establishment of blind refereeing for abstracts and journal articles would improve the chance for women economists to communicate with the profession.

We noted with regret this year the formation of an all-male editorial board for the new *Journal of Labor Economics*, published by the University of Chicago Press. At this writing, the editor has not given us the courtesy of a reply to our letter, sent last summer. Other journals also merit our attention in this regard.

Joan Robinson died in 1983, her prodigious accomplishments uncrowned by a Nobel Prize.

Shirley Kallek, Associate Director of the United States Census for Economic Fields, who was in charge of all of the work of the Bureau except that relating to population, also died this year. Among her other accomplishments was the organization of a section of the Bureau specializing in the economic analysis of microdata on business establishments. She was also Census liaison to the AEA Advisory Committee to the Census, a committee whose debates were instrumental in causing Census to end use of the term "head of household," to survey child support compliance, and to organize a conference on data needs for studying issues relating to women. A fellowship fund is being organized in her memory, and contributions to it may be made through CSWEP.

Another notable death this year was that of Beatrice N. Vaccara, who was Director in the Bureau of Industrial Economics of the Commerce Department. During the Carter Administration, she had served as Deputy Assistant Secretary for Domestic Economic Policy in the Treasury Department.

CSWEP Activities and Organization

CSWEP continued to debate this year how the organization could be most useful in furthering the recognition and prospects of women economists, whatever their specialty. The CSWEP sessions at the AEA and regional meetings tend to consist of papers concerning sex role issues in the economy and allied topics. While it is natural for CSWEP to have as one of its functions the furtherance of economic research on such

TABLE 5—DISTRIBUTION OF ACTIVITIES OF NEW PH.D. DEGREES BY SEX AND TYPE OF DEPARTMENT, ACADEMIC YEAR 1982–83

	All Ph.D. Depts.		Chair's Group		Other Ph.D. Depts.	
	No.	Percent	No.	Percent	No.	Percent
All Ph.D.s	772	100.0	353	100.0	419	100.0
Education	422	54.7	194	55.0	228	54.4
Government	67	8.7	31	8.8	36	8.6
Bus., Banking, Research	117	15.2	55	15.6	62	14.8
Int'l. Emp. Outside U.S.	113	14.6	51	14.4	62	14.8
Other	53	6.9	22	6.2	31	7.4
Male Ph.D.s	664	100.0	305	100.0	359	100.0
Education	357	53.8	164	53.8	193	53.8
Government	57	8.6	27	8.9	30	8.4
Bus., Banking, Research	103	15.5	48	15.7	55	15.3
Int'l. Emp. Outside U.S.	107	16.1	49	16.1	58	16.2
Other	40	6.0	17	5.6	23	6.4
Female Ph.D.s	108	100.0	48	100.0	60	100.0
Education	65	60.2	30	62.5	35	58.3
Government	10	9.3	4	8.3	6	10.0
Bus., Banking, Research	14	13.0	7	14.6	7	11.7
Int'l. Emp. Outside U.S.	6	5.6	2	4.2	4	6.7
Other	13	12.0	5	10.4	8	13.3

matters, some members have felt that a parallel way should be found to get exposure for women economists in other specialties.

In this regard, CSWEP is working to inform women economists of the mechanics of organizing sessions on the non-CSWEP part of the programs, and will be monitoring the degree of success women who attempt to do this meet with. Women economists who have made proposals to organize sessions at any meetings should inform the CSWEP Chair of the outcome.

We also continue to wrestle with ways to answer requests of prospective employers claiming to be looking for women candidates and asking us to help publicize their vacancies. Notices in the Newsletter are costly, and tend not to be timely. Moreover, the applications they encourage may be ignored. Lists of women who have faculty appointments currently, and lists of recent publications by women authors or coauthors are in process of compilation. Although these lists may prove useful, it is possible that other methods might prove worthwhile, and we continue to be on the lookout for them.

On the occasion of last spring's request for dues, we asked if members would like to volunteer for activities with CSWEP. We got a very encouraging response. A number of members will help out at the AEA convention, but we feel that there are many other possibilities which we have yet to organize or initiate. One possibility might be a clearinghouse for the provision of expertise for testimony before Congress and the State Legislatures, as well as in court proceedings. This would have to be done in a way consistent with AEA's nonpartisan and tax exempt status.

Committee W of the American Association of University Professors has sent letters to CSWEP and to all of the women's caucuses in the other academic professions, asking "what, if anything, is being done to review undergraduate texts and curricula for sex bias, and what is being done to introduce women's issues into the curriculum." In the coming year, CSWEP will consider how we might act to move this work forward in economics.

Nancy Ruggles has earned our sincere thanks for her supervision of computer work on the CSWEP membership list and the production of the CSWEP Roster. The Roster continues to provide an invaluable means of locating women economists by area and specialty. Ruggles is passing this work to Joan

TABLE 6—DISTRIBUTION OF PH.D. STUDENT SUPPORT, BY TYPE OF SUPPORT, SEX, AND DEPARTMENT, ACADEMIC YEAR 1982–83

	All Ph.D. Depts.		Chair's Group		Other Ph.D. Depts.	
	No.	Percent	No.	Percent	No.	Percent
All Students	7,248	100.0	3,254	100.0	3,994	100.0
Tuition Only	401	5.5	185	5.7	216	5.4
Stipend Only	560	7.7	220	6.8	340	8.5
Tuition + Stipend	3,333	46.0	1,506	46.3	1,827	45.7
No Support	2,034	28.1	948	29.1	1,086	27.2
No Record	920	12.7	395	12.1	525	13.1
Male Students	5,740	100.0	2,597	100.0	3,143	100.0
Tuition Only	306	5.3	141	5.4	165	5.6
Stipend Only	464	8.1	179	6.9	285	9.1
Tuition + Stipend	2,606	45.4	1,182	45.5	1,424	45.3
No Support	1,632	28.4	761	29.3	871	27.7
No Record	732	12.8	334	12.9	398	12.7
Female Students	1,508	100.0	657	100.0	851	100.0
Tuition Only	95	6.3	44	6.7	51	6.0
Stipend Only	96	6.4	41	6.2	55	6.5
Tuition + Stipend	727	48.2	324	49.3	403	47.4
No Support	402	26.7	187	28.5	215	25.3
No Record	188	12.5	61	9.3	127	14.9

Haworth, who has been one of CSWEP's most active and valued members. Also leaving the committee this year are Irma Adelman, Monique P. Garrity, and Janet C. Goulet, to whom much thanks are owed. Coming onto the committee will be Sharon Megdall of the University of Arizona-Phoenix, Lourdes Beneria of Rutgers University-New Brunswick, Bernadette Chachere of Hampton Institute, Michelle J. White of the University of Michigan, and Mary Fish of the University of Alabama.

BARBARA R. BERGMANN, *Chair*

Report of the Committee on U.S.–Soviet Exchanges

The Seventh U.S.–Soviet Economic Symposium was held at Yale University, June 4–6, 1983, on the subject "The Economics of Non-Renewable Resources." The atmosphere of the meeting was, as usual, informal and friendly. There was some improvement in the quality of the Soviet participation, compared with previous meetings held in the United States. The Soviet delegation was younger than usual, perhaps due partly to urging from this side, and it also included for the first time two women economists. The Soviet papers were reasonably well prepared, though not up to the quality of the U.S. presentations.

After the Symposium the Soviet delegation was scheduled to visit New York, Washington, and New Orleans. But, for reasons unknown to us, the Soviet authorities reduced the visit to one week from the usual two weeks. So New Orleans had to be cancelled, and time in New York and Washington was also curtailed.

The Eighth Symposium is scheduled to be held in the USSR in early June, 1984. The subject is "Structural Change in the U.S. and Soviet Economies," which we interpret as involving a fifty- to sixty-year time span. The meeting will probably be held in Riga or Tallinin, followed by visits to Moscow and other cities. Organization of the U.S. delegation is well under way, and it appears that there will be no difficulty in fielding a strong team.

LLOYD REYNOLDS, *Chair*

Report of the Committee on U.S.–China Exchanges

In 1983, while official exchanges in economics through the Committee on Scholarly Communication with the People's Republic of China (CSCPRC) were slow moving, many other channels of exchanges continued to be active.

Herbert Simon, Chairman of CSCPRC in charge of exchanges in social sciences and humanities, during his visit to China in the spring of 1983 to lecture and to conduct joint experimental research with Chinese psychologists, had discussions with officials of the Chinese Academy of Social Sciences (CASS) concerning exchanges in social sciences with CASS. In economics, in spite of the cancellation of the intended trip of a team of U.S. economists to study Chinese agricultural economic policy in 1982, CSCPRC expressed to the Chinese officials in CASS that we would welcome a delegation of Chinese economists to visit the United States to study the formulation of economic policy here. Under discussion was a delegation from CASS to pay a return visit to the United States following the visit of a group of economists from the National Bureau of Economic Research in 1982, whom CASS had received as an official delegation from the United States. This Chinese delegation, to be hosted by NBER and CSCPRC, did not arrive in 1983, but is scheduled for September 1984.

One problem facing the discussions between CSCPRC and CASS is the difference in the interests of these two organizations. CSCPRC insists on exchanges which will yield mutual benefits including the promotion of the research interests of American scholars working on China. CASS is more interested in cosponsoring lectures by American economists or workshops for the Chinese economists to learn modern economics, and is less interested in cooperative research. Besides this basic problem, cultural exchanges were terminated by PRC in April 1983 because of political and diplomatic events such as the granting of political asylum to the Chinese tennis player Hu Na.

A strong force promoting further exchanges in economics is the firm policy of the Chinese government to expand foreign trade and investment and to modernize the economy. In his report to the Twelfth National Congress of the Communist Party of China on September 1, 1982, the Party General Secretary Hu Yaobang stated, "We must improve our study and application of economics and scientific business management and continuously raise the level of economic planning and administration and the operation and management of enterprises and institutions." A manifestation of this policy is the series of short courses on project evaluation sponsored by the World Bank. In the Spring of 1983, Arnold C. Harberger and James Henderson participated in such a course in China.

Other examples include two series of workshops on agricultural economics sponsored by the Agricultural Development Council with financial support from the Ford Foundation. One series was jointly sponsored by the Chinese Association of Agricultural Sciences Societies. D. Gale Johnson conducted the workshop in the summer of 1983. The second series was cosponsored by the Chinese Academy of Agricultural Science. Besides these workshops, the Agricultural Development Council also sponsors visits by Chinese and American scholars, and provides fellowships to Chinese students to study agricultural economics.

The Chinese Ministry of Education has been supporting a small number of students to study economics abroad. In 1983 a number of American universities had graduate students from China studying economics, including Chicago, Cornell, Illinois, Minnesota, Rochester, Princeton, and Yale, among others. About half of these students were sponsored by the Chinese Ministry of Education. Many Chinese economics professors were visiting American universities, some being supported by the Luce Foundation. The Chinese Ministry of Education is committed to promoting the education of modern economics in China.

The communication of economic ideas between both sides of the Pacific continued in

1983. To facilitate this communication I completed a textbook *The Chinese Economy* in 1983, to be published by Harper & Row in the summer of 1984. This book is an attempt to apply the tools of economic analysis to study the Chinese economy. It may help American students to understand the Chinese economy and Chinese students to understand modern economics through its application to China.

By the end of 1983 it became very clear that further communications between American and Chinese students and scholars in economics would continue in the future.

GREGORY C. CHOW, *Chair*

New Books from The World Bank

WORLD TABLES, THE THIRD EDITION
VOLUME I: Economic Data
VOLUME II: Social Data

from the data files of the World Bank

The third edition of the WORLD TABLES continues the World Bank's policy of periodically making available the products of its ongoing collection, analysis, and updating of economic, demographic, and social data on most countries and territories of the world.

Like its predecessors, the new edition provides historical time series data for individual countries for the basic economic and social variables. Introduced in this edition is a set of tables on industrial statistics, trade in manufacturers classified by the International Standard Industrial Classification (ISIC), and the main results of the International Comparison Project (ICP) on purchasing-power parities and real gross domestic product. Also new is a set of tables that gives time series for the social and demographic data for each country.

VOLUME I	**$50.00** ***hardcover***	**$25.00** ***paperback***
VOLUME II	**$25.00** ***hardcover***	**$12.50** ***paperback***
The two-volume set	**$65.00** ***hardcover***	**$32.50** ***paperback***

Suitable for Course Use

COST-BENEFIT ANALYSIS
Issues and Methodologies

Anandarup Ray

Anandarup Ray examines the numerous important contributions to the theory and practice of cost-benefit analysis, consolidating much of the recent work in the area and focusing on aspects that continue to be controversial.

Among the topics discussed are alternative types of valuation functions, differential weighting for income inequality and for disparities in the consumption of basic needs, shadow exchange rates and the valuation of nontraded and traded goods and services, valuation of savings and budget constraints, and concepts of discount rates and of shadow rates.

$22.50 ***hardcover*** **$9.50** ***paperback*** June

Published for The World Bank by

THE JOHNS HOPKINS UNIVERSITY PRESS
Baltimore, Maryland 21218

Economics from The Free Press

MARKETS AND HIERARCHIES: ANALYSIS AND ANTITRUST IMPLICATIONS

A Study in the Economics of Internal Organization

OLIVER E. WILLIAMSON, *University of Pennsylvania*

"Oliver Williamson's work has brought to our attention in a fruitful study the ways that economic and other organizations differ in their behavior from the pure market model....I find his work especially important because in introducing a need for new principles in understanding internal organization, he has not lost sight of the importance of external market pressures in the direction of economic activity. The publication of this book is a powerful addition to our knowledge in this area and equally a stimulus to further research." —Kenneth J. Arrow, Nobel laureate in economics

286 pages paper 0-02-934780-7 $11.95

AN INTRODUCTION TO MONETARY THEORY AND POLICY *Third Edition*

DWAYNE WRIGHTSMAN, *University of New Hampshire*

Organized around monetary supply, demand, and the market equilibrating process, this revised and expanded edition explains the Depository Deregulation and Monetary Control Act of 1980, discusses the new official definitions of money (M-1, M-2, and M-3), presents an entirely new chapter on formulating monetary policy, and analyzes the rational expectations approach to conducting policy.

342 pages paper 0-02-935920-1 $9.95

SOCIAL SECURITY AND PENSIONS IN TRANSITION

Understanding the American Retirement System

BRUNO STEIN, *New York University*

Stein traces the history of income maintenance programs from their origins to the present, demonstrating that we have inherited not only the goals of the past—care for the aged, income redistribution, political stability, sound employment relations—but their outmoded mechanisms as well. "For those interested in an overview of the programs that comprise the U.S. retirement system, this book provides an excellent introduction."—Alicia H. Munnell, *The Bankers Magazine*

308 pages paper 0-02-930860-7 $11.95

CONCENTRATION, MERGERS, AND PUBLIC POLICY

YALE BROZEN, *University of Chicago*

Brozen draws on important empirical research to show that economic power is not becoming concentrated in the hands of a few corporations, that collusion, monopoly pricing, and inferior goods and services do not typify concentration, and that price rises in concentrated industries have been more moderate than in decentralized ones. "...highly readable and often entertaining. If its lessons are heeded, American industry will be free to seek productivity gains and to improve its competitive position in domestic and world markets."—Ernest Gellhorn, University of Virginia

425 pages 0-02-904270-4 $29.95

REAGANOMICS: MEANING, MEANS, AND ENDS

JOHN KENNETH GALBRAITH, *Emeritus, Harvard University, and* PAUL W. McCRACKEN, *University of Michigan*

The 24th volume in the Charles C. Moskowitz Memorial Lectures at New York University brings together Galbraith's and McCracken's contrasting views on Reaganomics at midterm, Galbraith arguing that our Age of Organization—governmental, corporate, and union—has left the market defunct, and McCracken offering qualified praise of the Reagan administration's control of spending, stable monetary policy, emphasis on improved productivity, and devotion to the liberal economic order.

95 pages 0-02-922890-5 $12.95

Examine any of these books FREE for 30 days. At the end of that period either honor our invoice or return the book(s) without further obligation. Or send check or money order now and Free Press pays postage and handling.

To place a VISA or MasterCard order call TOLL-FREE 800-257-5755. (In NJ, AK, or HI dial direct 609-461-6500.)

Fp

THE FREE PRESS
A Division of Macmillan, Inc.
866 THIRD AVENUE, NEW YORK, NY 10022
Att: Eileen Kennedy

ETHICS
DYING
BLACK-
MAIL
SELF-
CONTROL
DAY-
DREAMS
TERRORISM

SOUND INVESTMENTS FROM COLUMBIA AND NYU

THE IMPACT OF THE MODERN CORPORATION

Edited by Betty Bock, Harvey J. Goldschmid, Ira M. Millstein, and F.M. Scherer. Commissioned by The Center for Law and Economics at Columbia University, the volume gathers papers on specific aspects of corporation size in relation to worker alienation, corporate philanthropy, community impact, technological change, and political impact. *Government and Business Series, Eli Noam, General Editor.* 400 pp., $12.50 pa, $35.00 cl *(August)*

COMPETITION, MONOPOLY, AND DIFFERENTIAL PROFIT RATES

ON THE RELEVANCE OF THE CLASSICAL AND MARXIAN THEORIES OF PRODUCTION PRICES FOR MODERN INDUSTRIAL AND CORPORATE PRICING

Willi Semmler. 240 pp., ***A King's Crown Paperback,*** $12.00; $30.00 cl *(August)*

AN ECONOMIC HISTORY OF THE MIDDLE EAST AND NORTH AFRICA

Charles Issawi. Now in paperback. "This is an enormously interesting and well executed book."
—The Economist.
Columbia Economic History of the Modern World.
300 pp., $12.50 pa

GROWTH IN THE AMERICAN SOUTH

CHANGING REGIONAL EMPLOYMENT AND WAGE PATTERNS IN THE 1960s and 1970s

Robert J. Newman. Gathers together an impressive body of documentation on the extent and magnitude of the redistribution of industry employment towards the South and offers a number of well-argued and significant explanations for the trend. **NEW YORK UNIVERSITY PRESS.** 256 pp., $30.00

THE NATURE AND SIGNIFICANCE OF ECONOMIC SCIENCE

SECOND EDITION

Lord Robbins. Fifty years after its first appearance, the author has written a reassessment of the book in which he takes note of, and issue with, some of his critics. **NEW YORK UNIVERSITY PRESS.** 200 pp., $15.00 pa, $30.00 cl

MONETARIST, KEYNESIAN, AND NEW CLASSICAL ECONOMICS

Jerome L. Stein. "An excellent summary of the current controversies in macroeconomics."*—Choice.*
Now in paperback. **NEW YORK UNIVERSITY PRESS.**
238 pp., $12.50 pa

THE SUNBELT/SNOWBELT CONTROVERSY

THE WAR OVER FEDERAL FUNDS

Robert J. Dilger. Now available in paperback, this is the first book to study the growing influence of caucuses in the modern Congress. "Offers rare insights into governmental policy making."*—Library Journal.*
NEW YORK UNIVERSITY PRESS. 240 pp., $12.50 pa *(June)*

JAPAN'S EMERGING MULTINATIONALS

AN INTERNATIONAL COMPARISON OF POLICIES AND PRACTICES

Susumu Takamiya and Keith Thurley, Editors. This study of British, American and Japanese electronics firms evaluates worker perception and performance as the basis for discussing Japanese management of international business. **UNIVERSITY OF TOKYO PRESS.**
250 pp., $20.00 *(August)*

WEALTH AND MONEY IN THE ARISTOTELIAN TRADITION

A STUDY IN SCHOLASTIC ECONOMIC SOURCES

Odd Langholm. A study of the origins of economics which examines for the first time a number of unprinted commentaries on Aristotle, Nicole Oresme, and other philosophers. **UNIVERSITETSFORLAGET.**
112 pp., $22.00

OFFSHORE ADVENTURE

A PICTORAL HISTORY OF THE NORWEGIAN PETROLEUM INDUSTRY

Thorvald Buch Hansen. **UNIVERSITETSFORLAGET.** Illus.
64 pp., $44.00

Send check or money order to Dept. JN at the address below, including $2.00 per order for postage and handling.

New York University Press is Distributed Worldwide by

COLUMBIA UNIVERSITY PRESS

136 South Broadway, Irvington, NY 10533

Econometric/Statistical Package

xSTAT

Mainframe Capability, Accuracy and Speed for Your IBM Personal Computer

xSTAT: A Reliable, Economical Alternative to Mainframe Computing

- Is designed for the analysis of **VERY LARGE** as well as small data sets. It has no limit on sample size (total number of observations); and allows up to 50 variables per regression.
- Can generate and save an **X'X product matrix** the first time a data set is used. Subsequent regressions can be made from it to eliminate repetitive readings of input data normally required in statistical packages—very time-consuming when floppy disks are used for storing data.
- Supports Intel 8087 80-bit Math Co-processor for **FAST NUMBER CRUNCHINGS** (has software emulation for systems without the co-processor to attain same accuracy). Double precision is used in matrix operations.
- Produces estimated coefficients that are identical to mainframe TSP results to four or more significant digits.
- Executes at **HIGH SPEED** for IBM PC with a hard disk or a RAM drive—it takes only a fraction of a second to complete a regression run. For a system with floppy disks, it is as fast as the speed of the output media. The output can be directed to disk, printer or console.
- Works in either batch or prompt mode; requires no special command language. **EASY TO USE.**

THE PACKAGE INCLUDES:

Data Reformatting and Transformation Program

- Can reformat both fixed and free format datafile and is handy for processing data downloaded from mainframe. Can also reformat datafile into DIF (data interchange format) for input to Visicalc and the Lotus 1-2-3 Program.
- Allows selection of subsamples based on value range or sequential order. Converts fractions to decimal numbers (e.g., stock quotes, 32⅛ = > 32.125).
- Has 14 mathematical/logical transformation functions for vector multiplication, division, addition, subtraction, logarithm, exponent, lagged variables, first difference, dummy variables, and time trend.

Descriptive Statistics

- Minimum, maximum, mean, standard deviation, sum of squares, correlation coefficients, **CROSS-TABS**, etc.

Ordinary Least Squares (Multiple regressions) and Two-stage Least Squares

- Generates variance/covariance matrix of coefficients, standard error, t-statistic, R-square, adjusted R-square, Durbin-Watson statistic and F-statistic. Calculates predicted values, residuals, and produces residual plot.
- Allows subsample selection; can also estimate regressions from product moment matrix (contingency table).

Pooling Cross-Section and Time-Series Data

- Estimates regressions for samples with multi-period observations (e.g., panel study). Automatic stacking of data matrix (diagonal block matrix). Estimates coefficients and F-statistics for testing various hypotheses of (1) a common intercept and slope; (2) a common slope and different intercepts; (3) a common intercept and different slopes; (4) different intercepts and different slopes.

SYSTEM REQUIREMENTS

IBM PERSONAL COMPUTER (PC or XT), COMPAQ or IBM PC compatible systems: 1 floppy disk drive or a hard disk system; DOS 1.1 or 2.0 and 128K memory.

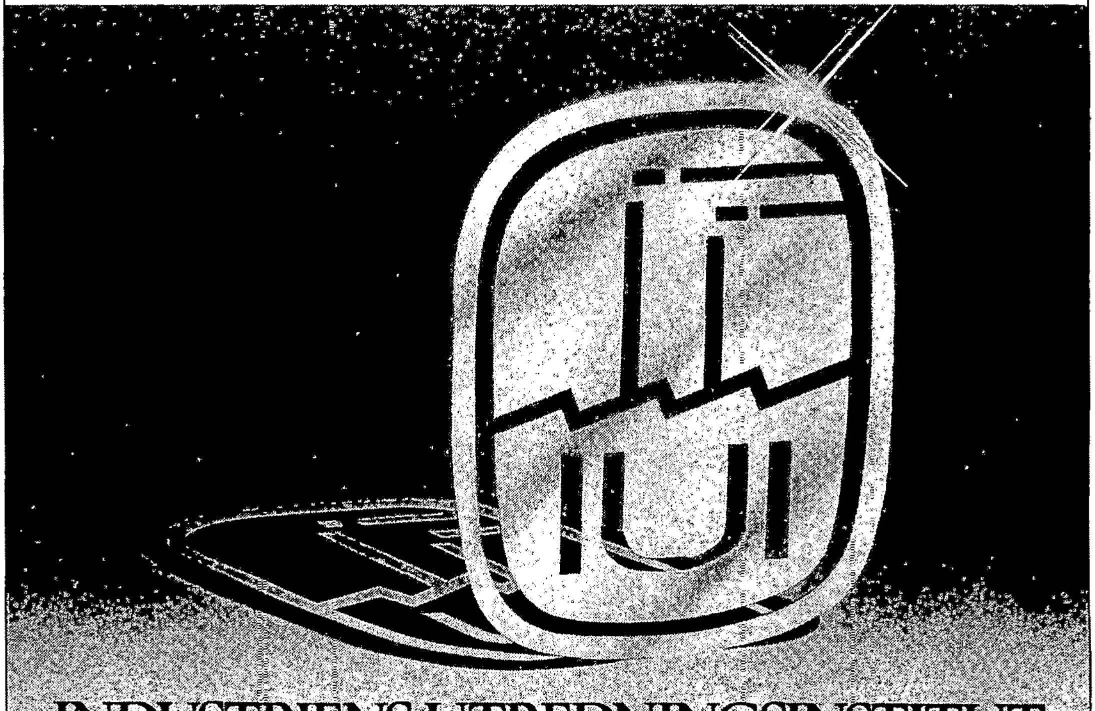
PRICE: $180 for the complete package.
(Send payment with order and save postage and handling. Accept Visa & MasterCard. Or order through your local dealer.)

RELATED PRODUCTS

For Apple II, II+, IIe:
DATA*TRANS: Telecommunications and Data Transfer
SampleCal: Automatic Sample Size Calculator
Exec Slides: Ordered sequence of graphic displays

For IBM Personal Computer:
PC Telecom: Telecommunications and Data Transfer

MING TELECOMPUTING INC.
Telecommunications and Statistics for Microcomputers
P.O. BOX 101
Lincoln Center, MA 01773
(617) 259-0391
(SAMPLE OUTPUTS ARE AVAILABLE UPON REQUEST.)



The American Economic Review

ARTICLES

JUNE 1984

THE AMERICAN erican Econom[illegible] ATION

•Published at George Banta Co., Inc. Menasha, Wisconsin. The publication number is ISSN 0002-8282.

•*THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the Directory, and Supplements, is published by the American Economic Association and is sent to all members five times a year: March; May; June; September; December.

Dues for 1984, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature*, are as follows:

$33.00 for regular members with rank of assistant professor or lower, or with annual income of $15,840, or less;

$39.60 for regular members with rank of associate professor, or with annual income of $15,840 to $26,400;

$46.20 for regular members with rank of full professor, or with annual income above $26,400;

$16.50 for junior members (registered students). Certification must be submitted yearly.

Institutional subscriptions (libraries, institutions, or firms) are $100.00 a year. Individual subscriptions are $65.00 a year. Only subscriptions to both publications will be accepted. Single copies of either journal may be purchased from the Secretary's office.

In countries other than the United States, add $9.20 to cover extra postage.

•Correspondence relating to the Directory, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

•Second-class postage paid at Nashville, Tennessee and at additional mailing offices, Printed in U.S.A.

•Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.

Founded in 1885

Officers

President
CHARLES L. SCHULTZE
The Brookings Institution

President-Elect
CHARLES P. KINDLEBERGER
Massachusetts Institute of Technology

Vice Presidents
ZVI GRILICHES
Harvard University
ROBERT L. HEILBRONER
New School for Social Research

Secretary
C. ELTON HINSHAW
Vanderbilt University

Treasurer
RENDIGS FELS
Vanderbilt University

Managing Editor of The American Economic Review
ROBERT W. CLOWER
University of California-Los Angeles

Managing Editor of The Journal of Economic Literature
MOSES ABRAMOVITZ
Stanford University

Executive Committee

Elected Members of the Executive Committee
ANN F. FRIEDLAENDER
Massachusetts Institute of Technology
JOSEPH E. STIGLITZ
Princeton University
WILLIAM D. NORDHAUS
Yale University
A. MICHAEL SPENCE
Harvard University
VICTOR R. FUCHS
Stanford University
JANET L. NORWOOD
Bureau of Labor Statistics

EX OFFICIO Members
GARDNER ACKLEY
The University of Michigan
W. ARTHUR LEWIS
Princeton University

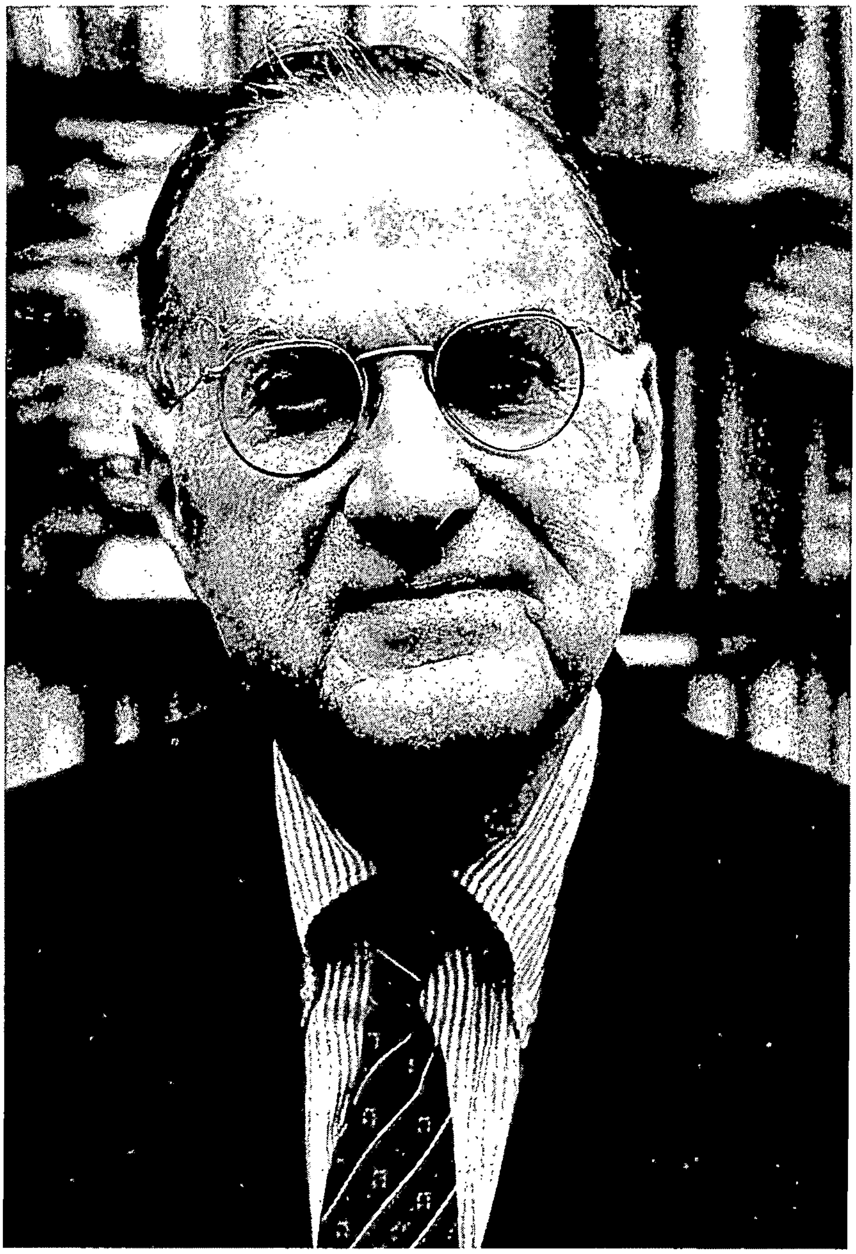
ABRAM BERGSON

DISTINGUISHED FELLOW

1983

Abram Bergson has been a leader of this generation of economists on two fronts. Bergson created the new welfare economics in one happy stroke, clarifying for us post-Paretians the purpose of achieving our Pareto optimality. The unifying concept of a Bergson social welfare function, by now a generic common noun, showed how ethics has a role in political economy—one that needs to be and can be kept distinguished from positivistic economic reality. The hedonistic utilitarianism of the old welfare economics was made understandable in its special place.

At the same time that Abram Bergson continued his theoretical researches, extending and deepening our understanding of consumer's surplus, cardinal utility, Bergson constant-elasticity-of-substitution functions, and more, he became the dean of scholars analyzing the Soviet Union. Bergson has inspired a corps of specialists on the economics of the centralized societies. His example of objective scholarship has elevated the intellectual climate of the entire discipline.



Abram Bergson

THE AMERICAN ECONOMIC REVIEW

ROBERT W. CLOWER
Managing Editor

JOHN G. RILEY
Associate Editor

WILMA ST. JOHN
Production Editor

THERESA DE MARIA
Assistant Editor

●Manuscripts and editorial correspondence relating to the regular quarterly issues of this *Review* and the *Papers and Proceedings* should be addressed to Robert W. Clower, Managing Editor, *AER* Editorial Office, University of California, Los Angeles, CA 90024. Manuscripts should be submitted in triplicate and in acceptable form, and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: $25 for members; $50 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

●No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

June 1984

VOLUME 74, NUMBER 3

Articles

Shorter Papers

Economic Theory in the Mathematical Mode

By GERARD DEBREU*

I

If a symbolic date were to be chosen for the birth of mathematical economics, our profession, in rare unanimous agreement, would select 1838, the year in which Augustin Cournot published his *Recherches sur les Principes Mathématiques de la Théorie des Richesses*. Students of the history of economic analysis could point out contributions made to mathematical economics as early as the beginning of the eighteenth century. They could also point out Johann Heinrich von Thünen's *Der Isolierte Staat*, 1826, a prototypical example of the use of mathematical reasoning in economic theory with little mathematical formalism. But Cournot stands out as the first great builder of mathematical models explaining economic phenomena. Among his successors in the nineteenth century and the early twentieth century, the highest prominence will be given in this lecture to Léon Walras (1834–1910), the founder of the mathematical theory of general economic equilibrium, to Francis Y. Edgeworth (1845–1926), and to Vilfredo Pareto (1848–1923). All three lived long enough into the twentieth century to have increased, for all Nobel Laureates, the value of the economics prize, had it, like the other prizes, been initiated in 1901.

If 1838 is the symbolic birthdate of mathematical economics, 1944 is the symbolic beginning of its contemporary period. In that year, John von Neumann and Oskar Morgenstern published the first edition of the *Theory of Games and Economic Behavior*, an event that announced a profound and extensive transformation of economic theory. In the following decade, powerful intellectual stimuli also came from many other directions. In addition to von Neumann and Morgenstern's book, Wassily Leontief's input-output analysis, Paul Samuelson's *Foundations of Economic Analysis*, Tjalling Koopmans' activity analysis of production, and George Dantzig's simplex algorithm were frequent topics of discussion, notably at the Cowles Commission when I joined it on June 1, 1950. To become associated at that time with a strongly interactive group which provided the optimal environment for the type of research that I wanted to do was an exceptional privilege.

One leading motivation for that research was the study of the theory of general economic equilibrium. Its goals were to make the theory rigorous, to generalize it, to simplify it, and to extend it in new directions. The execution of such a program required the solution of several problems in the theory of preferences, utility, and demand. It led to the introduction into economic theory of new analytical techniques borrowed from diverse fields of mathematics. Occasionally it made it necessary to find answers to purely mathematical questions. The number of research workers involved was, at first, small and slowly increasing, but in the early 1960's it began to grow more rapidly.

The most primitive of the concepts of the theory I will survey and discuss is that of the commodity space. One makes a list of all the commodities in the economy. Let l be their finite number. Having chosen a unit of measurement for each one of them, and a sign convention to distinguish inputs from outputs (for a consumer inputs are positive, outputs negative; for a producer inputs are

*University of California, Berkeley, CA 94720. This article is the lecture Gerard Debreu delivered in Stockholm, Sweden, December 8, 1983, when he received the Nobel Prize in Economic Sciences, The article is copyright © the Nobel Foundation. It is published here with the permission of the Nobel Foundation, and is included in the volume *Les Prix Nobel 1983*.

I thank Robert Anderson, Frank Hahn, Werner Hildenbrand, Herbert Scarf, Stephen Smale, and especially George and Helen Break, for many helpful comments.

negative, outputs positive), one can describe the action of an economic agent by a vector in the commodity space R^l. The fact that the commodity space has the structure of a real vector space is a basic reason for the success of the mathematization of economic theory. In particular convexity properties of sets in R^l, a recurring theme in the theory of general economic equilibrium, can be fully exploited. If, in addition, one chooses a unit of account, and if one specifies the price of each one of the l commodities, one defines a price-vector in R^l, a concept dual to that of a commodity-vector. The value of the commodity-vector z relative to the price-vector p is then the inner product $p \cdot z$.

One of the aims of the mathematical theory that Walras founded in 1874–77 is to explain the price-vector and the actions of the various agents observed in an economy in terms of an equilibrium resulting from the interaction of those agents through markets for commodities. In such an equilibrium, every producer maximizes his profit relative to the price-vector in his production set; every consumer satisfies his preferences in his consumption set under the budget constraint defined by the value of his endowment-vector and his share of the profits of the producers; and for every commodity, total demand equals total supply. Walras and his successors for six decades perceived that his theory would be vacuous without an argument in support of the existence of at least one equilibrium, and noted that in his model the number of equations equals the number of unknowns, an argument that cannot convince a mathematician. One must, however, immediately add that the mathematical tools that later made the solution of the existence problem possible did not exist when Walras wrote one of the greatest classics, if not the greatest, of our science. It was Abraham Wald, starting from Gustav Cassel's (1918) formulation of the Walrasian model, who eventually in Vienna in 1935–36 provided the first solution in a series of papers that attracted so little attention that the problem was not attacked again until the early 1950's.

Kenneth Arrow has told in his Nobel lecture (1974) about the path that he followed to the point where it joined mine. The route that led me to our collaboration was somewhat different. After having been influenced at the Ecole Normale Supérieure in the early 1940's by the axiomatic approach of N. Bourbaki to mathematics, I became interested in economics toward the end of World War II. The tradition of the School of Lausanne had been kept alive in France, notably by François Divisia and by Maurice Allais, and it was in Allais' formulation in *A la Recherche d'une Discipline Economique* (1943) that I first met, and was captivated by, the theory of general economic equilibrium. To somebody trained in the uncompromising rigor of Bourbaki, counting equations and unknowns in the Walrasian system could not be satisfactory, and the nagging question of existence was posed. But in the late 1940's several essential elements of the answer were still not readily available.

In the meantime, an easier problem was solved, and its solution contributed significantly to that of the existence problem. At the turn of the century, Pareto had given a characterization of an optimal state of an economy in terms of a price system, using the differential calculus. A long phase of development of Pareto's ideas in the same mathematical framework came to a resting point with the independent contributions of Oscar Lange (1942) and of Allais (1943). In the summer of 1950, Arrow, at the Second Berkeley Symposium on Mathematical Statistics and Probability, and I, at a meeting of the Econometric Society at Harvard, separately treated the same problem by means of the theory of convex sets. Two theorems are at the center of that area of welfare economics. The first asserts that if all the agents of an economy are in equilibrium relative to a given price-vector, the state of the economy is Pareto optimal. Its proof is one of the simplest in mathematical economics. The second provides a deeper economic insight and rests on a property of convex sets. It asserts that associated with a Pareto optimal state s of an economy, there is a price-vector p relative to which all the agents are in equilibrium (under conditions that, here as elsewhere, I cannot fully specify). Its proof is based on the observation that in the commodity space R^l, the a priori given en-

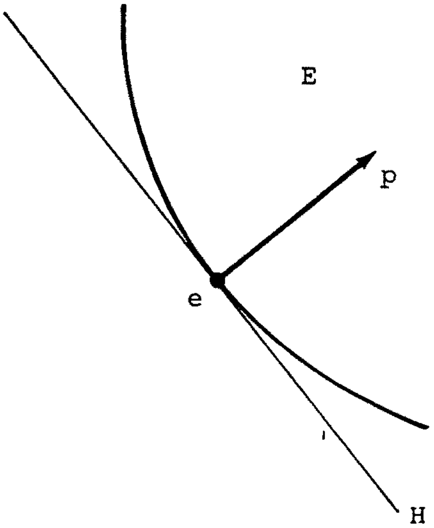


FIGURE 1

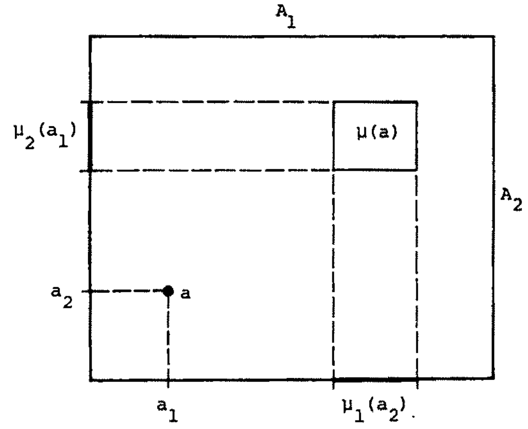


FIGURE 2

dowment-vector e of the economy is a boundary point of the set E of all the endowment-vectors with which it is possible to satisfy the preferences of all consumers at least as well as in the state s. Under conditions insuring that the set E is convex, there is a supporting hyperplane H for E through e. A vector p orthogonal to the hyperplane H, pointing towards E has all the required properties. (See Figure 1.) The treatment of the problem thus given by means of convexity theory was rigorous, more general and simpler than the treatment by means of the differential calculus that had been traditional since Pareto. The supporting hyperplane theorem (more generally the Hahn-Banach theorem, Debreu, 1954a) seemed to fit the economic problem perfectly. Especially relevant to my narrative is the fact that the restatement of welfare economics in set-theoretical terms forced a reexamination of several of the primitive concepts of the theory of general economic equilibrium. This was of great value for the solution of the existence problem.

In the year I joined the Cowles Commission, I learned about the Lemma in von Neumann's article of 1937 on growth theory that Shizuo Kakutani reformulated in 1941 as a fixed point theorem. I also learned about the applications of Kakutani's theorem made by John Nash in his one-page note of 1950 on "Equilibrium Points in N-Person Games" and by Morton Slater in his unpublished paper, also of 1950, on Lagrange multipliers. Again there was an ideal tool, this time Kakutani's theorem, for the proof that I gave in 1952 of the existence of a social equilibrium generalizing Nash's result. Since the transposition from the case of two agents to the case of n agents is immediate, we shall consider only the former which lends itself to a diagrammatic representation. Let the first agent choose an action a_1 in the a priori given set A_1, and the second agent choose an action a_2 in the a priori given set A_2. Knowing a_2, the first agent has a set $\mu_1(a_2)$ of equivalent reactions. Similarly, knowing a_1, the second agent has a set $\mu_2(a_1)$ of equivalent reactions. (See Figure 2.) $\mu_1(a_2)$ and $\mu_2(a_1)$ may be one-element sets, but in the important case of an economy with some producers operating under constant returns to scale, they will not be. The state $a = (a_1, a_2)$ is an equilibrium if and only if $a_1 \in \mu_1(a_2)$ and $a_2 \in \mu_2(a_1)$, that is, if and only if $a \in \mu(a) = \mu_1(a_2) \times \mu_2(a_1)$.

In other words, a is an equilibrium state if and only if it is a fixed point of the correspondence $a \mapsto \mu(a)$ from $A = A_1 \times A_2$ to A itself. Conditions insuring that Kakutani's theorem applies to A and μ guarantee the existence of an equilibrium state. In our article of 1954, Arrow and I cast a competitive economy in the form of a social system of the preceding type. The agents are the consumers, the producers, and a fictitious price setter. An appropriate definition of the set of reactions of the price setter to an excess demand vector makes the concept of equilibrium for that social system equivalent to the concept of competitive equilibrium for the original economy. In this manner a proof of existence, resting ultimately on Kakutani's theorem, was obtained for an equilibrium of an economy made up of interacting consumers and producers. In the early 1950's, the time had undoubtedly come for solutions of the existence problem. In addition to the work of Arrow and me, begun independently and completed jointly, Lionel McKenzie at Duke University proved the existence of an "Equilibrium in Graham's Model of World Trade and Other Competitive Systems" (1954), also using Kakutani's theorem. A different approach taken independently by David Gale (1955) in Copenhagen, Hukukane Nikaido (1956) in Tokyo, and Debreu (1956) in Chicago permitted the substantial simplification given in my *Theory of Value* (1959) of the complex proof of Arrow and Debreu.

Following that approach we consider a price-vector p different from 0 in R^l_+, the closed positive orthant of R^l. The reactions of the consumers and of the producers in the economy to p yield an excess demand vector z in R^l, whose coordinates represent for each commodity the (positive, zero, or negative) excess of demand over supply. Since the vector z may not be uniquely determined, one is led to introduce the set $Z(p)$ of the excess demand vectors associated with p, a set which is invariant if p is multiplied by a strictly positive real number. If every commodity in the economy can be freely disposed of, p^* is an equilibrium price-vector if and only if there is in $Z(p^*)$ a vector all of whose coordinates are negative or zero, that is, if and only if $Z(p^*)$ intersects R^l_-, the closed negative orthant of R^l. The fact that every consumer satisfies his budget constraint implies that all the points of $Z(p)$ are in or below the hyperplane through the origin of R^l orthogonal to p. (See Figure 3.) Additional conditions on Z suggested by Kakutani's theorem establish the existence of an equilibrium p^*.

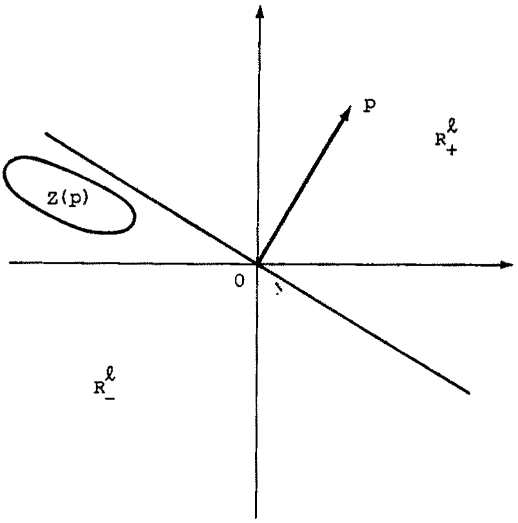


FIGURE 3

A proof of existence is now considered a necessary adjunct of a model proposing a concept of economic equilibrium, and in a recent survey (Debreu, 1982) more than 350 publications containing existence proofs of that type were listed. One of the most complex among these, because of the generality at which it aimed, was my article (1962).

During the past three decades, several other approaches to the problem of existence have been developed. Without attempting a systematic survey such as those prepared for Arrow and Intriligator (1981–84) by Stephen Smale (ch. 8), by Debreu (ch. 15), by E. Dierker (ch. 17), and by Herbert Scarf (ch. 21), one must explicitly mention two of them here.

Given an arbitrary strictly positive price-vector p, we now consider the case in which the reactions of the consumers and of the producers in the economy determine a unique excess demand vector $F(p)$. We also assume that the budget constraint of every consumer is exactly satisfied. Then one has

$$\text{Walras' Law} \qquad p \cdot F(p) = 0.$$

This equality suggests that the price-vector p be normalized by restricting it to the strictly positive part S of the unit sphere in R^l, for then the vector $F(p)$, being orthogonal to p, can be represented as being tangent to the

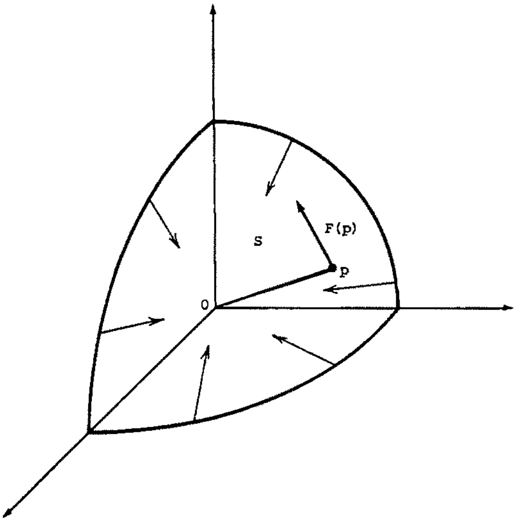


FIGURE 4

sphere S at p. (See Figure 4.) In mathematical terms, the excess demand function F defines a vector field on S. This representation turned out to be the key to the general characterization of excess demand functions that I will discuss later. It also provides an existence proof in the case of a boundary condition on F, meaning in economic terms that excess demand becomes large when some prices tend to zero, and in mathematical terms that the excess demand points inward near the boundary of S. For a continuous vector field, this property implies that there is at least one point p^* of S for which $F(p^*)=0$. This equality of demand and supply for every commodity expresses that p^* is an equilibrium price-vector.

The second approach concerns the development of efficient algorithms for the computation of approximate equilibria, an area of research in which Scarf (1973) played the leading role. The search for algorithms of that class is a natural part of the program of study of general economic equilibrium. Yet the decisive stimulus came unexpectedly from the solution of a problem in game theory, when C. E. Lemke and J. T. Howson (1964) provided an algorithm for the solution of two-person non-zero-sum games. The computation of equilibria has found its way into a large number of applications and has added an important new aspect to the theory of general economic equilibrium.

The explanation of equilibrium given by a model of the economy would be complete if the equilibrium were unique, and the search for satisfactory conditions guaranteeing uniqueness has been actively pursued (an excellent survey is found in Arrow and Hahn, 1971, ch. 9). However, the strength of the conditions that were proposed made it clear by the late 1960's that global uniqueness was too demanding a requirement and that one would have to be satisfied with local uniqueness. Actually, that property of an economy could not be guaranteed even under strong assumptions about the characteristics of the economic agents. But one can prove, as I did in 1970, that, under suitable conditions, in the set of all economies, the set of economies that do *not* have a set of locally unique equilibria is negligible. The exact meaning of the terms I have just used and the basic mathematical result on which the proof of the preceding assertion rests can be found in Sard's theorem to which Stephen Smale introduced me in conversations in the summer of 1968. The different parts of the solution fell into place at Milford Sound on the South Island of New Zealand. On the afternoon of July 9, 1969, when my wife Françoise and I arrived, intermittent rain and overcast weather that dulled the view tempted me to work once more on what had become a long tantalizing problem, and, this time, ideas quickly crystallized. The next morning a cloudless sky revealed the Sound in its midwinter splendor.

The "suitable conditions" to which I alluded are differentiability conditions which, in the present situation, are essentially unavoidable. As for the term "negligible," it means, in the case of a finite-dimensional set of economies, "contained in a closed set of Lebesgue measure zero." The main ideas of the proof can be conveyed intuitively in the simple case of an exchange economy with m consumers. The demand function f_i of the ith consumer associates with every pair (p, w_i) of a strictly positive price-vector p and a positive wealth (or income) w_i the demand $f_i(p, w_i)$ in the closed positive orthant R^l_+ of the commodity space. The ith consumer is characterized by his demand function f_i and by his endowment-vector e_i in the strictly positive orthant P of R^l. The functions f_i are kept fixed and assumed to be continuously differentiable. Therefore, the economy is described by the list $e=(e_1,\dots,e_m)$ of the m endowment-vectors in P^m. The price-vector p being restricted to belong to S, the strictly positive part of the unit sphere, the excess demand vector associated with a pair (p,e) in $S\times P^m$ is

$$F(p,e)=\sum_{i=1}^{m}\left[f_i(p,p\cdot e_i)-e_i\right].$$

The equilibrium manifold M (Smale, 1974; Balasko, 1975) is the subset of $S\times P^m$ defined by $F(p,e)=0$, an equality which, because of Walras' Law, imposes only $l-1$ constraints. Under the assumptions made, M is a differentiable manifold and its dimension is $\dim M=\dim P^m+\dim S-(l-1)=lm=\dim P^m$. Now let T be the projection from M into P^m, and define a critical economy e as an economy such that it is the projection of a point (e,p) of M where the Jacobian of T is singular, geometrically where the tangent linear manifold of dimension lm does not project onto P^m. (See Figure 5.) By Sard's theorem the set of critical economies is closed and of Lebesgue measure zero. A regular economy, outside the negligible critical set, not only has a discrete set of equilibria; it also has a neighborhood in which the set of equilibria varies continuously as a function of the parameters defining the economy. The study of regular economies thus forms a basis for the analysis of the determinateness of equilibrium and of the stability of economic systems. Moreover, the continuity of the set of equilibria in a neighborhood of a regular economy insures that the explanation of equilibrium provided by the model is robust with respect to unavoidable errors in the measurement of the parameters. Once again, a mathematical result, Sard's theorem, was found to fit exactly the needs of economic theory. The study of regular economies has been an active research area in the last decade, and Smale, Balasko, and Andreu Mas-Colell (1984) are among its main contributors.

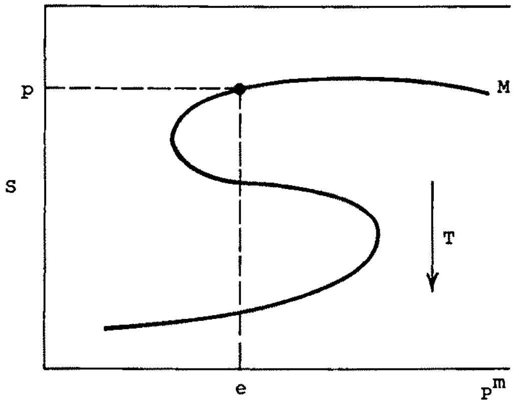


FIGURE 5

Departing from chronological order, I now return to the late 1950's and to the early 1960's, and to the beginning of the theory of the core of an economy. Edgeworth (1881) had given a persuasive argument in support of the common imprecise belief that markets become more competitive as the number of their agents increases in such a way that each one of them tends to become negligible. He had specifically shown that his "contract-curve" tends to the set of competitive equilibria in a two-commodity economy with equal numbers of consumers of each one of two types. His brilliant contribution stimulated no further work until Martin Shubik (1959) linked Edgeworth's contract curve with the game theoretical concept of the core (D. B. Gillies, 1953). The first extension of Edgeworth's result was obtained by Scarf (1962), and the complete generalization to the case of an arbitrary number of commodities and of types of consumers was given by Debreu and Scarf (1963). Associated with our joint paper is one of my most vivid memories of the instant when a problem is solved. Scarf, then at Stanford, had met me at the San Francisco airport in December 1961, and as he was driving to Palo Alto on the freeway, one of us, in one sentence, provided a key to the solution; the other, also in one sentence, immediately provided the other key; and the lock clicked open. Once again, the basic mathematical result was the supporting hyperplane theorem for convex sets. The theorem that we had proved remained special, because it applied only to economies with a given number of types of consumers and an equal, increasing number of consumers of each type. Generalizations were soon forthcoming. Robert Aumann (1964) introduced the concept of an atomless measure space of economic agents, a natural mathematical formulation of the concept of an economy with a large number of agents, all of them negligible. Under notably weak conditions, Aumann proved that for such an economy the core coincides with the set of competitive equilibria. Karl Vind (1964) then pointed out that the proper mathematical tool for the proof of that striking result was Lyapunov's (1940) theorem on the convexity, and compactness, of the range of an atomless finite-dimensional vector measure. Out of these contributions grew an extensive literature that included among its high points Yakar Kannai's (1970) and Truman Bewley's (1973) articles, and that culminated in Werner Hildenbrand's book (1974). This was surveyed recently in Arrow and Intriligator (1982) by Hildenbrand (ch. 18).

In a different direction, a formalization of an economy with a large number of negligible agents was proposed by Donald Brown and Abraham Robinson (1972), who introduced the sophisticated techniques of Nonstandard Analysis in economic theory. Remarkably, this approach eventually led to the elementary inequalities of Robert Anderson (1978) on the extent of competitiveness of allocations in the core in an economy with a finite number of agents.

In the mid-1970's, the theory of the core and the theory of regular economies were joined in the study of the rate of convergence of the core to the set of competitive equilibria. Lloyd Shapley (1975) had shown that convergence could be arbitrarily slow. Debreu (1975) then proved that in the case of increasing equal numbers of agents of each of a finite number of types, the rate of convergence to the set of competitive equilibria of a *regular* economy is of the same order as the reciprocal of the number of agents. The extension of this result from replicated economies to more general sequences of economies was provided by Birgit Grodal (1975).

Intimately linked with the contemporary development of the theory of general economic equilibrium was that of the theory of preferences, utility, and demand. New results in the latter were in some cases required, in others motivated by the former. The primitive concepts in the theory of preferences of a consumer are his consumption set X, a subset of R^l, and his preference relation $\lesssim$, a complete preorder on X. We shall say that a real-valued function u on X is a utility function if it represents the preference relation $\lesssim$ in the sense that

$$[x \lesssim y] \Leftrightarrow [u(x) \leqq u(y)].$$

A necessary and sufficient condition for the existence of a continuous utility function is that the set $G = \{(x, y) \in X \times X | x \lesssim y\}$ be closed relative to $X \times X$ (Debreu, 1954b; 1964). Although more abstract than the familiar concept of an infinite family of indifference sets in R^l, the concept of a single set G in $R^l \times R^l$ is far simpler as two more instances illustrate.

To say that an agent has preferences similar to that of another means for a mathematical economist that a topology has been introduced on the set of preferences. This was done by Kannai (1970), in an article whose publication was long delayed. The prospect of comparing two preference relations $\lesssim$ and $\lesssim'$ on the two consumption sets X and X' (now assumed to be closed) is daunting if one thinks of each preference relation as an infinite family of indifference sets in R^l. It becomes appealing if one thinks of each preference relation as a closed subset of $R^l \times R^l$ (Debreu, 1969). The topology on the set of preferences was at the basis of the theory of the core in Hildenbrand (1974). It was indispensable for the work that Kannai (1974) and Mas-Colell (1974) did on the approximation of a convex preference relation by convex preference relations representable by concave utility functions.

The other instance pertains to preference relations representable by differentiable utility functions. The traditional approach, by focusing on the consumption set X in R^l, raised delicate integrability questions (extensively surveyed by Leonid Hurwicz, 1971, ch.

9). In contrast, a differentiable preference relation $\lesssim$ can simply be defined by the condition that the boundary of the associated set G is a differentiable manifold in $R^l \times R^l$ (Debreu, 1972).

In all these developments, the theory of preferences was stimulated and helped by questions asked about the utility function u such as "When is u continuous?," "When is u concave?," "When is u differentiable?" Yet another instance is provided by the study of a preference relation $\lesssim$ defined on the product X of n sets $X_1, \ldots, X_n$. The question now is whether the preference relation $\lesssim$ can be represented by a utility function of the form

$$u(x)=\sum_{i=1}^{n} u_i(x_i),$$

where x is the n-list $(x_1, \ldots, x_n)$ and for every i, $x_i \in X_i$. This problem was studied by Leontief (1947a,b) and by Samuelson (1947, ch. 7), by means of the differential calculus. It can be studied by topological methods (Debreu, 1960) which bring out more clearly the essential independence property on which the solution is based.

The last example from the theory of preferences, utility, and demand will be the problem of the characterization of the excess demand function of an economy. We consider an exchange economy $\mathscr{E}$ with m consumers. As before, the demand function f_i of the ith consumer associates with a pair (p, w_i) of a price-vector p in the strictly positive part S of the unit sphere in R^l and of a wealth (or income) w_i in the set R_+ of nonnegative real numbers, a consumption vector $f_i(p, w_i)$ in the closed positive orthant R^l_+ of R^l. If the ith consumer has a preference relation $\lesssim_i$ on R^l_+, then $f_i(p, w_i)$ is a commodity-vector that satisfies $\lesssim_i$ under the budget constraint $p \cdot z \leqq w_i$. The economy $\mathscr{E}$ is defined by specifying for the ith consumer $(i=1,\ldots,m)$ the demand function f_i and the endowment-vector e_i in R^l_+. The aggregate excess demand function of the economy is the function F defined by

$$\text{(a)} \qquad F(p)=\sum_{i=1}^{m}\left[f_i(p, p \cdot e_i)-e_i\right].$$

Under weak standard assumptions, the function F (1) is continuous and (2) satisfies Walras' Law. Hugo Sonnenschein (1972, 1973) asked whether these two properties characterize F. Specifically, given F satisfying (1) and (2), can one find m consumers with demand functions f_i and endowment-vectors e_i satisfying (a)? Sonnenschein conjectured that the answer was affirmative and made the first attack on this problem. Rolf Mantel (1974) proved Sonnenschein's conjecture in the case of continuously differentiable demand functions, and Debreu (1974) in the general case. The proof appearing in this last article was inspired by, and rests on, the representation of the excess demand function F as a vector-field on the strictly positive part of the unit sphere. The characterization of aggregate excess demand functions so obtained has several applications. It shows that the hypothesis of preference satisfaction (or equivalently of utility maximization) puts essentially no restriction on F, that a theorem on the existence of a general economic equilibrium is equivalent to a fixed point theorem (via an observation of Hirofumi Uzawa, 1962), and that any dynamic behavior can be observed for an economy operating under a tâtonnement process (as the examples of global instability of Scarf, 1960, presaged). One impact of that characterization has been the redirection of research on aggregate demand functions toward a specification of the distribution of the characteristics of the economic agents. The first theoretical result explaining the "Law of Demand" (Hildenbrand, 1983) was a product of that redirected research.

II

Having surveyed in some detail, as tradition requires, the work cited by the Royal Swedish Academy of Sciences, I turn to issues of methodology in economic theory.

Contemporary developments in the theory of general economic equilibrium took Walras' work as their point of departure, but some of Walras' ideas had a long lineage that included Adam Smith's (1776) profound insight. Smith's idea that the many agents of an economy, making independent decisions, do not create utter chaos but actually contribute to producing a social optimum, raises indeed a scientific question of central importance. Attempts to answer it have stimulated the study of several of the problems that every economic system must solve, such as the efficiency of resource allocation, the decentralization of decisions, the incentives of decision makers, the treatment of information.

In the past few decades, that wide range of problems has been the subject of an axiomatic analysis in which primitive concepts are chosen, assumptions concerning them are formulated, and conclusions are derived from those assumptions by means of mathematical reasoning disconnected from any intended interpretation of the primitive concepts. The benefits of the axiomatization of economic theory have been numerous. Making the assumptions of a theory entirely explicit permits a sounder judgment about the extent to which it applies to a particular situation. Axiomatization may also give ready answers to new questions when a novel interpretation of primitive concepts is discovered. As an illustration, consider the concept of a commodity, which had meant traditionally a good or a service whose physical properties and whose delivery date and location are specified. In the case of an uncertain environment, Arrow (1953) added to those characteristics of a commodity the event in which delivery will take place. In this manner one obtains, without any change in the form of the model, a theory of uncertainty in which all the results of the theory of certainty are available (Debreu, 1959, ch. 7). Axiomatization, by insisting on mathematical rigor, has repeatedly led economists to a deeper understanding of the problems they were studying, and to the use of mathematical techniques that fitted those problems better. It has established secure bases from which exploration could start in new directions. It has freed researchers from the necessity of questioning the work of their predecessors in every detail. Rigor undoubtedly fulfills an intellectual need of many contemporary economic theorists, who therefore seek it for its own sake, but it is also an attribute of a theory that is an effective thinking tool. Two other major attributes of an effective theory are simplicity and generality. Again, their aesthetic appeal suffices to make them desirable ends in themselves for the designer of a theory. But their value to the scientific community goes far beyond aesthetics. Simplicity makes a theory usable by a great number of research workers. Generality makes it applicable to a broad class of problems.

In yet another manner, the axiomatization of economic theory has helped its practitioners by making available to them the superbly efficient language of mathematics. It has permitted them to communicate with each other, and to think, with a great economy of means. At the same time, the dialogue between economists and mathematicians has become more intense. The example of a mathematician of the first magnitude like John von Neumann devoting a significant fraction of his research to economic problems has not been unique. Simultaneously, economic theory has begun to influence mathematics. Among the clearest instances are Kakutani's theorem, the theory of integration of correspondences (Hildenbrand, 1974), algorithms for the computation of approximate fixed points (Scarf's ch. 21 in Arrow and Intriligator, 1981–84), and of approximate solutions of systems of equations (Smale's ch. 8 in Arrow and Intriligator, 1981).

III

In narratives of their careers, scientists try to acknowledge the main influences to which they responded, and the support they received from other scientists and from different institutions, even though such attempts are unlikely to be entirely successful. To all the persons and organizations I have named, I want to add the outstanding education system I have known in France, and the Centre National de la Recherche Scientifique which made my conversion from mathematics to economics possible. After my move to the United States in 1950, I was associated with three great universities (Chicago, Yale, and Berkeley) where scientific research is a natural way of life; and during the last two decades the Economics Program of the National Science Foundation has given me, more than anything else, time for that research. All those institutions have provided a superb environment for the task that had to be performed.

REFERENCES

Allais, M., *A la Recherche d'une Discipline Economique*, Paris: Imprimerie Nationale, 1943.

Anderson, R. M., "An Elementary Core Equivalence Theorem," *Econometrica*, November 1978, *46*, 1483–87.

Arrow, K. J., "An Extension of the Basic Theorems of Classical Welfare Economics," in J. Neyman, ed., *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press, 1951, 507–32.

______, "Le Rôle des Valeurs Boursières pour la Répartition la Meilleure des Risques," *Económétrie*, Paris: Centre National de la Recherche Scientifique, 1953, 41–48.

______, "General Economic Equilibrium: Purpose, Analytic Techniques, Collective Choice," *American Economic Review*, June 1974, *64*, 253–72.

______ **and Debreu, G.,** "Existence of an Equilibrium for a Competitive Economy," *Econometrica*, July 1954, *22*, 265–90.

______ **and Hahn, F. H.,** *General Competitive Analysis*, San Francisco: Holden Day, 1971.

______ **and Intriligator, M. D.,** *Handbook of Mathematical Economics*, Vols. I, II, III, Amsterdam: North-Holland, 1981–84.

Aumann, R. J., "Markets with a Continuum of Traders," *Econometrica*, January-April, 1964, *32*, 39–50.

Balasko, Y., "On the Graph of the Walras Correspondence," *Econometrica*, September-November 1975, *43*, 907–12.

Bewley, T. F., "Edgeworth's Conjecture," *Econometrica*, September-November 1973, *41*, 425–54.

Brown, D. J. and Robinson, A., "A Limit Theorem on the Cores of Large Standard Exchange Economies," *Proceedings of the National Academy of Sciences of the U.S.A.*, 1972, *69*, 1258–60.

Cassel, K. G., *Theoretische Sozialökonomie*, Leipzig: C. F. Winter, 1918.

Cournot, A., *Recherches sur les Principes Mathèmatiques de la Thèorie des Richesses*, Paris: L. Hachette, 1838.

Dantzig, G. B., "Maximization of a Linear Function of Variables Subject to Linear Inequalities," in T. C. Koopmans, ed., *Activity Analysis of Production and Allocation*, New York: Wiley & Sons, 1951, 339–47.

Debreu, G., "The Coefficient of Resource Utilization," *Econometrica*, July 1951, *19*, 273–92.

______, "A Social Equilibrium Existence Theorem," *Proceedings of the National Academy of Sciences*, 1952, *38*, 886–93.

______, (1954a) "Valuation Equilibrium and Pareto Optimum," *Proceedings of the National Academy of Sciences*, 1954, *40*, 588–92.

______, (1954b) "Representation of a Preference Ordering by a Numerical Function," in R. M. Thrall et al., eds., *Decision Processes*, New York: Wiley & Sons, 1954, 159–65.

______, "Market Equilibrium," *Proceedings of the National Academy of Sciences*, 1956, *42*, 876–78.

______, *Theory of Value, An Axiomatic Analysis of Economic Equilibrium*, New York: Wiley & Sons, 1959.

______, "Topological Methods in Cardinal Utility Theory," in: K. J. Arrow et al., eds., *Mathematical Methods in the Social Sciences, 1959*, Stanford: Stanford University Press, 1960, 16–26.

______, "New Concepts and Techniques for Equilibrium Analysis," *International Economic Review*, September 1962, *3*, 257–73.

______, "Continuity Properties of Paretian Utility," *International Economic Review*, September 1964, *5*, 285–93.

______, "Neighboring Economic Agents," in *La Décision*, Colloques Internationaux du Centre National de la Recherche Scientifique No. 171, Paris, 1969, 85–90.

______, "Economies with a Finite Set of Equilibria," *Econometrica*, May 1970, *38*, 387–92.

______, "Smooth Preferences," *Econo-*

metrica, July 1972, *40*, 603–15; "Smooth Preferences: A Corrigendum," *Econometrica*, July 1976, *44*, 831–32.

______, "Excess Demand Functions," *Journal of Mathematical Economics*, 1974, *1*, 15–21.

______, "The Rate of Convergence of the Core of an Economy," *Journal of Mathematical Economics*, 1975, *2*, 1–7.

______, "Regular Differentiable Economies," *American Economic Review Proceedings*, May 1976, *66*, 280–87.

______, "Existence of Competitive Equilibrium," in K. J. Arrow and M. D. Intriligator, eds., *Handbook of Mathematical Economics*, Vol. II, Amsterdam: North-Holland, 1982, ch. 15.

______ **and Scarf, H.,** "A Limit Theorem on the Core of an Economy," *International Economic Review*, September 1963, *4*, 235–46.

______ **and Koopmans, T. C.**, "Additively Decomposed Quasiconvex Functions," *Mathematical Programming*, 1982, *24*, 1–38.

Dierker, E., "Regular Economies," in K. J. Arrow and M. D. Intriligator, eds., *Handbook of Mathematical Economics*, Vol. II, Amsterdam: North-Holland, 1982, ch. 17.

Divisia, F., *Economique Rationnelle*, Paris: Doin, 1928.

Edgeworth, F. Y., *Mathematical Psychics*, London: Kegan Paul, 1881.

Gale, D., "The Law of Supply and Demand," *Mathematica Scandinavica*, 1955, *3*, 155–69.

Gillies, D. B., "Some Theorems on *n*-Person Games," unpublished doctoral dissertation, Princeton University, 1953.

Grodal, B., "The Rate of Convergence of the Core for a Purely Competitive Sequence of Economies," *Journal of Mathematical Economics*, 1975, *2*, 171–86.

Hildenbrand, W., *Core and Equilibria of a Large Economy*, Princeton: Princeton University Press, 1974.

______, "Core of an Economy," in K. J. Arrow and M. D. Intriligator, eds., *Handbook of Mathematical Economics*, Vol. II, Amsterdam: North-Holland, 1982, ch. 18.

______, "On the 'Law of Demand'," *Econometrica*, July 1983, *51*, 997–1019.

Hurwicz, L., "On the Problem of Integrability of Demand Functions," in J. S. Chipman et al., eds., *Preferences, Utility, and Demand*, New York: Harcourt Brace Jovanovich, 1971, 174–214.

Kakutani, S., "A Generalization of Brouwer's Fixed Point Theorem," *Duke Mathematical Journal*, 1941, *8*, 457–59.

Kannai, Y., "Continuity Properties of the Core of a Market," *Econometrica*, November 1970, *38*, 791–815.

______, "Approximation of Convex Preferences," *Journal of Mathematical Economics*, 1974, *1*, 101–06.

Koopmans, T. C., *Activity Analysis of Production and Allocation*, New York: Wiley & Sons, 1951.

Lange, O., "The Foundations of Welfare Economics," *Econometrica*, July–October 1942, *10*, 215–28.

Lemke, C. E. and Howson, J. T., Jr., "Equilibrium Points of Bimatrix Games," *Journal of the Society of Industrial and Applied Mathematics*, 1964, *12*, 413–23.

Leontief, W. W., *The Structure of the American Economy, 1919–1929*, Cambridge: Harvard University Press, 1941.

______, (1947a) "A Note on the Interrelation of Subsets of Independent Variables of a Continuous Function with Continuous First Derivatives," *Bulletin of the American Mathematical Society*, 1947, *53*, 343–50.

______, (1947b) "Introduction to a Theory of the Internal Structure of Functional Relationships," *Econometrica*, October 1974, *15*, 361–73.

Lyapunov, A. A., "On Completely Additive Vector-Functions," *Izvestija Akademii Nauk SSSR*, 1940, *4*, 465–78.

McKenzie, L. W., "On Equilibrium in Graham's Model of World Trade and Other Competitive Systems," *Econometrica*, April 1954, *22*, 147–61.

Mantel, R., "On the Characterization of Aggregate Excess Demand," *Journal of Economic Theory*, March 1974, *7*, 348–53.

Mas-Colell, A., "Continuous and Smooth Consumers: Approximation Theorems," *Journal of Economic Theory*, July 1974, *8*,

305–36.

______, *The Theory of General Economic Equilibrium; a Differentiable Approach*, Cambridge: Cambridge University Press, 1984.

Nash, J. F., "Equilibrium Points in *N*-Person Games," *Proceedings of the National Academy of Sciences of the U.S.A.*, 1950, *36*, 48–49.

Neumann, J. von, "Über ein Ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes," *Ergebnisse eines Mathematischen Kolloquiums*, 1937, *8*, 73–83.

______ **and Morgenstern, O.**, *Theory of Games and Economic Behavior*, Princeton: Princeton University Press, 1944.

Nikaido, H., "On the Classical Multilateral Exchange Problem," *Metroeconomica*, August 1956, *8*, 135–45.

Pareto, V., *Manuel d'Economie Politique*, Paris: Giard, 1909.

Samuelson, P. A., *Foundations of Economic Analysis*, Cambridge: Harvard University Press, 1947.

Sard, A., "The Measure of the Critical Points of Differentiable Maps," *Bulletin of the American Mathematical Society*, 1942, *48*, 883–90.

Scarf, H., "Some Examples of Global Instability of Competitive Equilibrium," *International Economic Review*, September 1960, *1*, 157–72.

______, "An Analysis of Markets with a Large Number of Participants," *Recent Advances in Game Theory*, The Princeton University Conference, 1962.

______, (with the collaboration of T. Hansen), *The Computation of Economic Equilibria*, New Haven: Yale University Press, 1973.

______, "The Computation of Equilibrium Prices: An Exposition," in K. J. Arrow and M. D. Intriligator, eds., *Handbook of Mathematical Economics*, Vol. II, Amsterdam: North-Holland, 1982, ch. 21.

Shapley, L. S., "An Example of a Slow-converging Core," *International Economic Review*, June 1975, *16*, 345–51.

Shubik, M., "Edgeworth Market Games," *Contributions to the Theory of Games*, Vol. IV, *Annals of Mathematical Studies*, *40*, Princeton University Press, 1959.

Slater, M., "Lagrange Multipliers Revisited," Cowles Commission Discussion Paper, Mathematics 403, 1950.

Smale, S., "Global Analysis and Economics," IIA, *Journal of Mathematical Economics*, 1974, *1*, 1–14.

______, "Global Analysis and Economics," in K. J. Arrow and M. D. Intriligator, eds., *Handbook of Mathematical Economics*, Vol. I, Amsterdam: North-Holland, 1981, ch. 8.

Smith, A., *An Inquiry into the Nature and Causes of the Wealth of Nations*, 2 vols., London: W. P. Strahan and T. Cadell, 1776.

Sonnenschein, H., "Market Excess Demand Functions," *Econometrica*, May 1972, *40*, 549–63.

______, "Do Walras' Identity and Continuity Characterize the Class of Community Excess Demand Functions?," *Journal of Economic Theory*, August 1973, *6*, 345–54.

Thünen, J. H. von, *Der Isolierte Staat*, Hamburg: F. Perthes, 1826.

Uzawa, H., "Walras' Existence Theorem and Brouwer's Fixed Point Theorem," *Economic Studies Quarterly*, 1962, *8*, 59–62.

Vind, K., "Edgeworth-allocations in an Exchange Economy with Many Traders," *International Economic Review*, May 1964, *5*, 165–77.

Wald, A., "Über die Eindeutige Positive Lösbarkeit der Neuen Produktionsgleichungen," *Ergebnisse eines Mathematischen Kolloquiums*, 1935, *6*, 12–20.

______, (1936a) "Über die Produktionsgleichungen der Ökonomischen Wertlehre," *Ergebnisse eines Mathematischen Kolloquiums*, 1936, *7*, 1–6.

______, (1936b) "Über einige Gleichungs-Systeme der Mathematischen Ökonomie," *Zeitschrift für Nationalökonomie*, 1936, *7*, 637–70.

Walras, L., *Éléments d'Economie Politique Pure*, Lausanne: L. Corbaz and Company, 1874–77.

Capture and Ideology in the Economic Theory of Politics

By JOSEPH P. KALT AND MARK A. ZUPAN*

The economic theory of regulation long ago put public interest theories of politics to rest. These theories have correctly been viewed as normative wishings, rather than explanations of real world phenomena. They have been replaced by models of political behavior that are consistent with the rest of microeconomics (Anthony Downs, 1957; James Buchanan and Gordon Tullock, 1965; George Stigler, 1971; Sam Peltzman, 1976). Recently, however, debate has arisen over whether some version of a public interest theory of regulation will have to be readmitted to our thinking about actions and results in the political arena. What is at issue is the empirical importance of the altruistic, publicly interested goals of rational actors in determining legislative and regulatory outcomes (James Kau and Paul Rubin, 1979; Kalt, 1981; Peltzman, 1982).

This study assesses the nature and significance of publicly interested objectives in a particular instance of economic policymaking: U.S. Senate voting on coal strip-mining regulations. The existence of such objectives is, of course, no contradiction of the economic view of human behavior (Kenneth Arrow, 1972; Gary Becker, 1974); and may well be rooted in genetic-biological history (Becker, 1976; Jack Hirshleifer, 1978). Generally, however, individuals' altruistic, publicly interested goals have been given little attention. This reflects the judgment that such goals are so empirically unimportant as to allow the use of Occam's razor in positive models, or well-founded apprehensions that these goals are unusually difficult to identify, measure, and analyze. Notwithstanding the latter problem, we find that approaches which confine themselves to a view of political actors as narrowly egocentric maximizers explain and predict legislative outcomes poorly. The tracking and dissecting of the determinants of voting on coal strip-mining policy suggest that the economic theory of politics has been prematurely closed to a broader conception of political behavior.

I. Interests and Ideology in the Economics of Politics

A. *The Setting*: *Coal Strip-Mining Regulation*

The Surface Mining Control and Reclamation Act (SMCRA) was the product of a protracted political struggle. Congress twice passed versions of SMCRA—in 1974 and 1975—only to have them vetoed by President Ford. SMCRA was finally signed into law by President Carter on August 3, 1977. The Act requires the restoration of strip-mined land to its premining state. In addition, the Act established an Abandoned Mine Reclamation Fund and clarified previously indefinite property rights to water and land in areas underlain by strippable coal.

The Act reduces the use of environmental inputs and raises the costs of strip mining. This tends to raise the price of coal and generates income transfers *from* surface coal producers and coal consumers *to* underground producers and the consumers of environmental amenities. The combined losses of surface producers and coal consumers appear to be on the order of $1.4 billion per year (split approximately 70/30, respectively; Kalt, 1983). After accounting for a small deadweight loss, the annual gains of noncoal environmental users and under-

*Departments of Economics, Harvard University, Cambridge, MA 02138, and Massachusetts Institute of Technology, Cambridge, MA 02139, respectively. We thank Harold Demsetz, Allen Jacobs, Paul Joskow, Thomas Romer, Richard Schmalensee, Harry Watson, Mark Watson, and workshop participants at Harvard, MIT, and the University of Chicago for helpful comments and suggestions. Peter Martin, Kevin Mohan, and Margaret Walls provided valuable research assistance. The support of the Sloan Foundation and the Energy and Environmental Policy Center at the Kennedy School of Government has been greatly appreciated.

ground producers are in the range of $1.3 billion (split roughly 90/10).

While incidence analysis can produce more or less precise estimates of particular market participants' gains or losses from SMCRA, there is clearly no reason to expect these economic stakes to translate one-for-one into political clout (Mancur Olson, 1971; Stigler, 1971). Indeed, one of the tasks of the economic theory of regulation and the research below is to describe how economic stakes map into political influence. For our purposes, incidence analysis indicates the *direction* of relevant parties' interests in SMCRA: the regulation of the environmental damage attendant to strip mining should be expected to be opposed by surface coal producers and coal consumers; underground coal producers and consumers of affected environmental amenities should support SMCRA.

B. *Related Research*

The tenor of economic ("capture") theories of regulation, when applied to a specific case such as SMCRA, might suggest that the incidence of the legislation summarizes not only the economics, but also the politics of the issue: narrowly self-interested underground coal producers and environmental consumers captured policymakers at the expense of narrowly self-interested coal consumers and surface coal producers. This line of reasoning, however, cannot be disproved. Since every economic policy decision produces transfers of wealth, it is always possible to infallibly relate political outcomes to distributional impacts. This approach, in fact, leaves open the question of whether the behavior and results we observe in the political arena are the product of something more than the parochial pecuniary interests of affected parties.

Probably the most basic proposition of economic, capture models of regulation is the (sometimes implicit) assertion that the altruistic, publicly interested goals of individuals are such insignificant factors in political processes that they are empirically uninteresting and dispensable. Stigler (1972) has noted the possibility of altruistic motives in political action. These might take the form of a sense of "civic duty," that is, a duty to serve the interests of the public. Pursuit of such a duty is a consumption activity that yields utility in the form of the warm glow of moral rectitude. Classifying this type of argument in the utility function as a "consumption motive" (as distinguished from the self-interested "investment motive" of increasing one's own wealth), Stigler asserts with respect to economic theories of politics: "The investment motive is rich in empirical implications, and the consumption motive is less well-endowed, so we should see how far we can carry the former analysis before we add the latter" (p. 104).

The sentiment of this assertion may yet prove to be supportable. A number of recent investigations, however, have suggested that policymakers' self-defined notions of the "public interest" are dominant explanatory factors in congressional voting behavior (Edward Mitchell, 1979; Kau and Rubin; Kalt, 1981). These studies have attempted to explain voting records on specific issues (for example, oil price controls) as functions of relevant economic interest variables *plus* some measure of the "ideological" orientation of congressmen. The latter is typically based on rating scales provided by ideological watchdog organizations such as the Americans for Democratic Action (ADA) or the Americans for Constitutional Action (ACA). The consistent findings are that economic interest variables play surprisingly weak roles in legislative outcomes, while the hypothesis of no ideological effect is quite easily rejected.

Peltzman (1982) has taken a critical look at these findings. The interests of constituents and the ideological "preferences" of their representatives are plausibly interrelated—perhaps with causation running from the former to the latter. The apparent importance of ideology may, therefore, be due to left-out economic interest variables. By examining Senate voting across a broad sample of issues, Peltzman is able to "explain away" most of the importance of ideological preferences with an extended array of constituent interest measures (for example, demographic characteristics). The research strategy behind these results, however, differs in a fundamental way from the approach taken in the research it critiques. Specifically, Peltzman examines a sample covering essentially the entire package of votes offered by senators to their constituents, rather than voting on a specific issue. The query remains whether conclusions reached at such a high level of bundling can safely be applied to the specific case. The economic theory of regulation has generally been put forth and applied as an issue-specific theory (Stigler, 1971; Peltzman, 1976; Burton Abrams and Russell Settle, 1978).

C. *Possible Sources of Ideological Voting*

In the jargon of recent research, the purported social objectives of political actors have been termed "ideology." Political ideologies are more or less consistent sets of normative statements as to best or preferred states of the world. Such statements are moralistic and altruistic in the sense that they are held as applicable to everyone, rather than merely to the actor making the statements. Accordingly, political ideologies are taken here to be statements about how government can best serve their proponents' conceptions of the public interest. Behavior in accord with such statements has two possible sources: 1) the direct appearance of altruism in actors' preference functions (termed "pure" ideology); and 2) a convenient signalling mechanism when information on political decisions is otherwise costly.[1]

1. *Pure Ideology in Voters and their Representatives*. Pure ideology, if it exists at all, is the manifestation of altruism in the political sector. The returns from the furtherance of an ideology appear to come in at least two forms. First, the successful promotion of an ideology may give individuals the satisfaction of knowing that they have concretely improved the lot of others. Second, even if the pursuit of ideology has no effect on others, individuals may derive satisfaction from "having done the right thing" (Stigler's consumption motive).[2] Do individuals really get utility from these sources? It is not our intention here to dispute tastes. We take the presence of ideological tastes as given by introspection and observation. Following Becker (1974), we also take the pursuit of such tastes to be rational—to be responsive, that is, to opportunity costs. This contrasts with the unfortunate terminology which characterizes altruistic-ideological behavior as "non-economic" (Peltzman, 1982) and/or "irrational" (Yoram Barzel and Eugene Silberberg, 1973). The rationality of ideological behavior is tested below.

Political behavior based on pure ideology may arise from either the publicly interested objectives of constituents or the independent publicly interested objectives of their representatives.

Constituents: Voters' ideological goals might include, for example, anticommunism, communism, Jeffersonian agrarianism, Rawlsian egalitarianism, and so on.[3] The presence of such goals poses no problems for the economic theory of politics. Publicly interested ideologues are just another special interest capable of capturing the political process, subject to the comparative statics of organizational costs and benefits as modeled by Peltzman (1976). We suspect, however, that most economists would conclude that the pursuit of ideological objectives is not an important phenomenon. At least on the basis of behavior observed in the market sector, this would appear to be well-founded. Is there any reason to expect pure ideological actions to be relatively more common in the

[1] These two sources of ideology have often been noted. Downs' exposition is particularly clear. Jerome Rothenberg (1965), Bruno Frey and Lawrence Lau (1968), and Albert Breton (1974) provide systematic theoretical treatments of pure ideology; and empirical implementation of the concept has been pushed furthest by the literature on the "paradox of voting" (William Riker and Peter Ordeshook, 1968; Robert Tollison and Thomas Willett, 1975; and Orley Ashenfelter and Stanley Kelley, 1975).

[2] Of course, in either case, individuals may also receive returns in the form of the esteem of, or even reciprocal favors from, other individuals (Becker, 1976).

[3] The notion of "public interest" embodied in any particular ideology need not include the promotion of economic efficiency (Tullock, 1982). Indeed, ideologies appear to typically center around the "equity" (i.e., rights and distributional assignments) side of the economists' equity-efficiency dichotomy.

political arena? Several factors suggest the answer may be affirmative.

First, altruistic ideological interests that depend upon actually improving the welfare of others have clear collective good attributes. The apparatus of government provides the classic Samuelsonian (1954) means (i.e., coercive power) for overcoming the free-rider problems that can plague a marketplace. Indeed, this apparatus may be made comparatively inexpensive for the representative altruist to the extent it can be hijacked and used to require outsiders to finance the benefits delivered to the altruist's targeted group.

Second, in much political activity, the individual has no meaningful prospect of influencing outcomes. In the case of large-number majoritarian elections, for example, the individual voter is generally incapable of promoting his or her investment interests. This observation has led to the recognition that altruistic-ideological preferences play central roles in motivating the act of voting. Nevertheless, it has typically been assumed that, once the decision to vote has been made, we can explain the ballot cast by reference to the voter's economic interests. As Geoffrey Brennan and Buchanan (1982) have pointed out, however, this is a *non sequitur*: if the decision to vote is based on consumption motives, it does not follow that these motives are set aside upon entering the voting booth. Comparing the consumption choices made in voting with investment decisions in the marketplace (emphasis in original):

> ...we may presume that the individual *cares* as to which outcome emerges from the voting process. But this does not permit us to presume that his choice in the polling booth *reflects* or corresponds with his preferences over outcomes. For the voter is not *choosing* between outcomes.... When the voter pulls a particular lever, the opportunity cost of doing so is not a particular policy forgone.... [pp. 14–15]
>
> [Thus]...the choice of candidate...depends overwhelmingly on tastes for showing "preferences" as such—and hardly at all on the evaluation of outcomes. Voting behavior is then to be understood perhaps as "symbolic" or "liturgical"...and [is] hardly at all like the choice among alternative investments. [p. 18]

Third, even when political participation is motivated by the prospect of pecuniary gains, such gains are often subject to substantial public goods problems. Pecuniary political gains commonly must be shared with large numbers of congruent parties (for example, all coal consumers or all environmental users). While private sector investments can be accompanied by free-rider problems, such problems are virtually the rule at the legislative level in U.S.-style democracy (the Chrysler and Lockheed cases notwithstanding). In contrast, at least that part of ideology based on individuals' tastes for the warm glow of moral rectitude is a strict private good in both the public and private sectors. Thus the opportunity cost (i.e., forgone pecuniary return) of ideology might be expected to be generally lower in the political arena than in the marketplace. Accordingly, the rational actor in the political sector would be expected to reveal behavior tilted relatively more toward altruistic-ideological objectives.

Representatives: Institutional attributes of the political sector may allow pure ideological action by representatives themselves. This opportunity could arise because, analogous to the case of management in the private corporation, there may be some separation of "ownership" by constituents and "control" by policymakers. Any such slack in the principal-agent relationship can be expected to result in policymaker independence or "shirking"—as Armen Alchian and Harold Demsetz (1972) call it.

Models of the specific-issue legislative process that have grown out of the economic theory of regulation (Peltzman, 1976) seldom leave room for policymaker shirking.[4] By

[4] Notable exceptions in which ideological shirking appears are the formal models of Rothenberg and Frey and Lau, although the focus of these models is primarily on policymakers' electoral success. In a model motivated by institutional attributes similar to those considered here and below, Robert Barro (1973) allows slack in the principal-agent relationship and (albeit, nonideological) utility maximization by elected officials.

endowing legislators with goals such as vote maximization, rather than own-utility maximization, such models preclude behavior that is not directly controlled by constituents. This conception of the strength of the principal-agent bond in the legislative process, however, does not seem to be in line with the conception of this bond that comes out of the property rights theory of institutions (Alchian and Demsetz; Michael Jensen and William Meckling, 1976) and the bulk of associated empirical evidence (see the summary by Louis De Alessi, 1982). Conditions under which the market system's invisible hand is likely to encounter difficulty in narrowing the separation of ownership and control would appear to be especially prevalent in the legislative context.[5]

First, the "market for control" (Henry Manne, 1965) is characterized by significant indivisibilities that impair adjustment at the margin by constituents. The typical constituent is presented with all-or-nothing choices between a small number of large bundles of issues to be addressed by policymakers over their tenure; *and* the market meets only infrequently—every six years in the case of the U.S. Senate. Second, as "hirers" of political representation, voter-owners have poor incentives to be well-informed. As Olson has stressed, collective decisions are subject to classic free-rider problems that affect participants' willingness to invest in the acquisition of information. Third, these free-rider problems are exacerbated by the fact that "ownership" by constituents is held under attenuated property titles. Political ownership is nontransferable, and, as residual claimants to the net benefits of correct decisions, constituent-owners promoting such decisions cannot easily capture resulting gains (see Alchian and Demsetz). Fourth, the political market is apparently subject to less than perfect competition (John Ferejohn, 1977). The provision of representation services in the U.S. political system takes place under conditions of effective duopoly; barriers to entry are significant (Abrams and Settle); elements of natural monopoly are present (Stigler, 1971); and collusion to prevent Tiebout-type (1956) competition is officially sanctioned. Finally, these attributes of the market for legislative seats create conditions conducive to "opportunism" in Oliver Williamson's (1975) sense; and the range of enforceable contractual agreements of the type examined by Benjamin Klein, Robert Crawford, and Alchian (1978) that might be struck to minimize opportunism is notably limited (for example, to the verbal agreement that "I will keep my campaign promises").

It must be stressed that none of this implies that shirking is costless to legislative representatives. Analogous to the position of shirking private sector managers vis-à-vis shareholders, representatives face some control through the voting booth, as well as more continuous pressure from constituents who have some ability to affect the pleasantness of the policymaker's working day, future employment opportunities, and other aspects of the returns to positions of policy responsibility. Nevertheless, legislative institutions such as the U.S. Senate would appear to be archetypical Alchian-Demsetz organizations in which agents are imperfectly policed by their principals—where "imperfectly" is defined relative to a nirvana world of zero policing costs. The implied result is some amount of own-welfare maximization by representatives at the expense of their constituents—an amount that may be optimal for constituents given the real world policing costs they face.

Any shirking by imperfectly policed representatives can be expected to center around those activities that have low opportunity costs (for example, in terms of reelection prospects) and/or poor substitutes off the job. Paralleling Becker's (1957) analysis of private managers' on-the-job consumption and Alchian and Reuben Kessel's (1962) examination of nonprofit institutions, shirking by legislators may focus on nonpecuniary perquisites of office holding—although opportunities for personal pecuniary gain are certainly available. The perquisites of political office range from "fact finding" junkets and postservice employment connections with rent-seeking interest groups to public notoriety, prestige, and the ability to use the

[5] Ideological shirking does not require that separation of ownership and control be *more* prevalent in the legislative setting than in the marketplace. It is sufficient that there be *some* principal-agent slack.

power of government to impose one's own pet theories of the "good" society. The last of these emoluments is almost uniquely available in the political sector and is what we have termed ideological consumption.

For a number of reasons, shirking in an activity such as Senate voting might be expected to have an ideological component. Morris Fiorina and Roger Noll (1978) have noted, for instance, that legislators' fates depend heavily on the provision to constituents of so-called "facilitation services" (for example, supportive intervention at other levels of government), as distinguished from their provision of floor votes. This is complemented by the fact that nonideological shirking on floor votes (for example, taking bribes, failing to be informed, or missing roll calls in favor of office parties) is comparatively costly as a result of institutional penalties, while a legislator does not face expulsion or censure for voting his or her "conscience" (i.e., ideology).[6] Furthermore, to the extent the individual legislator can rationally take the fate on the floor of any particular piece of legislation as given, that legislator's vote becomes valueless to any constituent—that vote, that is, has no impact on the economic well-being of constituents since it does not affect outcomes.[7] The only remaining value of the vote to the legislator, then, would be its consumption value—no constituent would be willing to pay anything for it. This is, of course, the legislative-floor analogue to Brennan and Buchanan's analysis (noted above) of voting booth behavior by citizens, although it would be inappropriate to conclude that the investment value of a vote in a place like the U.S. Senate is typically nil as it is in the very-large-number majoritarian voting booth. Lastly, the governmental apparatus is the preeminent mechanism for affecting broad social change. The opportunities this creates, if coupled with comparatively low costs to ideological shirking, would imply a self-selection process that attracts individuals with relatively intense demands for ideology to the political sector.

2. *Impure Ideology and Costly Information.* It is certainly possible that policymakers base their decisions on consultation with the precepts of an ideology when nothing more than narrow self-interest is being served. In a world in which information on the concordance between constituent interests and the consequences of policy proposals is scarce, political representatives may serve their investment motives (for example, the desire to get reelected) by relying on the dictates of an ideology as a shortcut to the service of their constituents' goals (Downs; Buchanan and Tullock). In this view, ideology plays the same role in the economic theory of the political process that managerial rules of thumb play in the theory of the profit-maximizing firm. The implication that the apparent ideologies of representatives are in fact proxies for constituents' interests suggests collinearity between measures of ideology and those interests. This implication has not received support in studies of voting on individual issues, but has been borne out in Peltzman's (1982) examination of voting on the aggregated bundle of issues addressed by senators. These apparently conflicting results are analyzed below within the context of a Downsian view of representative democracy.

Downs' seminal look at representative democracy suggests an important implication of ideology as a device for economizing on information: if legislators are not perfectly policed on every vote, the rational constituent could be expected to support representatives whose demands for pure ideology are intense relative to other motives for shirking. To be sure, as each specific issue arises between elections, the constituent prefers that representatives vote the constituent's interests on that issue, not their own ideologies. But the constituent faced with 1) an uncertain bundle of issues to be decided by representatives over their terms of office, 2) uncertainty about the effects of policy deci-

[6] Note, also, that shirking in the pecuniary form of taking bribes (at the expense of support maximization) is service to some constituent's interest and would show up accordingly in the empirical analysis of voting below, unless bribery is uncorrelated with measures of constituent influence. In the latter situation, bribery and other forms of nonideological shirking (say, failing to be informed on issues) should show up as white noise.

[7] This point is also noted, in the context of a parliamentary system, by Frey and Lau (p. 358).

sions, and 3) positive policing costs and hence shirking, can attempt to wind up on net on the winning sides of issues (in a pecuniary *or* nonpecuniary sense) by supporting candidates with appropriate ideologies. Upon election, such candidates will engage in pure ideological shirking on particular issues rather than permit themselves to be captured by the issue-specific interests that organize and present themselves at any particular moment; an economic incidence approach to issue-specific political economy will be inadequate. Over the full slate of issues, however, representatives' voting should fall in line with general indicators of their constituents' ideological and investment interests—as Peltzman (1982) finds. We now examine these Downsian implications in the context of the specific issue of SMCRA.

II. Senate Voting on SMCRA

A. *Study Design*

Our objective is to untangle the causal forces behind Senate voting on strip-mining controls. We seek to separate the effects of constituents' interests (economic and ideological) and senators' ideology. In addition, we would like to be able to uncover that part, if any, of senators' ideology which is purely publicly interested shirking and that part of ideology which merely stands in for otherwise difficult to identify constituent interests.

Voting on strip-mining legislation is observed as either a "yea" or a "nay." To measure senators' positions, a variable *ANTISTRIP* is constructed to reflect the frequency, f_i, with which the ith senator casts a vote unfavorable to strip mining. A senator voting an anti-strip position on r_i out of n_i opportunities has $f_i = r_i/n_i$. This frequency is bounded by zero and unity. Adjusting for $r_i = 0$, $r_i = n_i$, and heteroskedasticity (John Gart and James Zweifel, 1967), the weighted logit technique of Arnold Zellner and Tong Lee (1965) is employed in our econometric analysis. Thus

$$(1) \quad ANTISTRIP_i = \ln[(r_i + .5)/(n_i - r_i + .5)];$$

and has variance estimator:

$$(2) \quad \text{Var}_i = 1/(r_i + .5) + 1/(n_i - r_i + .5);$$

The variable *ANTISTRIP* is based on the voting of the 100 senators that served in the 95th Congress (1977–78); and is derived from 21 roll call votes in which the interests of surface coal producers, underground producers, coal consumers, and the consumers of environmental amenities were clearly delineated. These votes deal with either SMCRA or its vetoed predecessors.[8] Measures that would have raised the costs of surface mining were taken to be detrimental to surface mining—and conversely. Selected votes and their economic implications are not identical; and Charles Phelps (1982) suggests the possibility of aggregation problems in *ANTISTRIP*. In this case, however, results are unaffected when individual votes are used as dependent variables (see our earlier paper).

B. *Interests and their Influence*

The economic theory of regulation provides the basis for measures of constituents' interests and influence. Specifically, the *interests* constituents have in capturing the political process are their prospective gains or losses from any policy proposal. The *ability* constituents have to capture the political process depends critically on their ability to overcome the free-rider effects inherent in collective decisions. Any group's influence will depend positively on members' per capita stakes and the concentration of their interests; and negatively on the heterogeneity of members' objectives and group size. Where data permit, we employ variables reflecting determinants of group effectiveness, as well as the magnitude of groups' interests in SMCRA.

Turning first to the magnitude of groups' interests, we introduce variables reflecting the stakes of surface coal producers, underground coal producers, coal consumers, and

[8] Votes are taken from Congressional Quarterly, Inc., *Congressional Quarterly Almanac* (1973, 1975, 1977). A descriptive list is available in our 1983 paper.

consumers of affected environmental amenities.

Coal Producers: The variables *SURFRES* and *UNDERRES* measure each state's reserves (in Btus) in 1977 of surface and underground coal, respectively; and are expressed as fractions of state personal income to scale for the relative importance of coal production to states' economies. Reserve-based measures are used to proxy for the present value of SMCRA's impacts on coal resources. Particularly in many western states where strip mining was in its infancy in 1977, current production figures inadequately capture the present-valued importance of the industry to states' economies. Results of most interest, namely the relative roles of economic and ideological variables, are not affected by switching to production- or employment-based measures (see our earlier paper).[9]

There are significant differences across states in SMCRA's impact on strip mining costs. The variable *MC* measures the regulation-induced increase in the long-run average cost of surface mining in each state, as derived by ICF, Inc. (1977). Because surface mining interests were adversely affected by SMCRA, *MC* as well as *SURFRES* are expected to be negatively related to *ANTISTRIP*. The variable *UNDERRES* is expected to have a positive impact on *ANTISTRIP*.

Coal Consumers: The variable *CONSUME* is employed to represent the importance of coal consumption in each state. Electric power generation accounts for 78 percent of U.S. coal demand and *CONSUME* is the share of state electricity generated from coal in 1977.[10] It is preferred to other measures such as total coal Btus consumed per capita if it is primarily electric utilities, small in number and large in size, who overcome the free-rider problems that plague political lobbying.[11] The organizational effectiveness of the electric power industry is examined below. Reflecting the effect of SMCRA on coal prices, *CONSUME* is expected to be negatively related to *ANTISTRIP*.

Environmental Consumers: Environmental interests may be classed into two broad types: environmental users in the literal sense; and those for whom environmental protection represents an ideological cause. Empirical evidence on the existence of the latter is provided by William Schulze et al. (1981). They find that, based on willingness to pay, the most significant value of an undeveloped environment is derived from individuals' demands for just knowing that an area is used "properly," independent of whether such individuals ever visit or even plan to visit the area themselves. These values are notably altruistic-ideological. They arise from prescriptive opinions about what environmental uses are consistent with self-defined standards of ethical propriety and the public interest. These standards include the view of wilderness as an antidote for purported psychological costs of urbanization; the conception of the American West as a peculiar cultural and natural history lesson; the quasi-religious question of the propriety of appropriating of environmental resources for human ends; the social desirability of rapid economic growth; and the appropriate beneficiaries of public lands. Moreover, the results of Schulze et al. suggest that people's willingness to pay to uphold these precepts has standard comparative static properties—ideological environmentalism is just another economic good.

To capture the ideological interests that constituents have in SMCRA, we employ a variable *ENVIROS*. This is defined as state membership in the six largest environmental groups (as a fraction of voting-age population). Interestingly, the correlation between *ENVIROS* and measures of actual recreational use of the environment (for example, hunting and fishing, budgets for parks

[9] Reserves are from National Coal Association, *Coal Data 1978* (1979). These are highly correlated with the projections of new mine development over 1978–87 reported in McGraw-Hill, Inc., *Keystone Coal Industry Manual* (1977).

[10] *CONSUME* is from *Coal Data 1978*. The correlation between *CONSUME* and an alternative measure, total coal Btus consumed per capita, is 0.72. Results reported below are essentially unchanged when the alternative measure is utilized (see our earlier paper).

[11] Of the over 50 coal consumers who appeared before congressional hearings on SMCRA, only one was not a utility.

and recreation, visits to parklands) is quite low.[12] If senators' voting has been captured by ideological environmentalists, *ENVIROS* should have a positive effect on *ANTISTRIP*.

The interests of actual nonmining consumers of the environment threatened by strip mining are represented by three variables: *HUNTFISH, SPLITRIGHTS*, and *UNRECLAIMED*. The variable *HUNTFISH* is defined as the number of hunting and fishing licenses as a percentage of state population. It is highly correlated with other measures of outdoor recreational activity. The analysis in Kalt (1983) indicates that outdoor recreation is little threatened by strip mining; and *HUNTFISH* is consistently insignificant in the econometric analysis. Since results of interest are invariant with respect to the inclusion of *HUNTFISH*, the variable is excluded here (see our earlier paper). The variable *SPLITRIGHTS* captures the support of ranchers, farmers, lumberers, and other noncoal business interests for legislation that preserved their preferential rights to the large land areas underlaid by federally controlled strippable coal. The economic values at stake are measured by the agriculture/timber revenue yield of the disputed surface acres, expressed as a percentage of state personal income. Similarly, *UNRECLAIMED* measures the prospective value of already stripped but unrestored acres to noncoal interests who stood to benefit from the Abandoned Mine Reclamation Funds' subsidies. The variables *SPLITRIGHTS* and *UNRECLAIMED* should be positively related to *ANTISTRIP*.[13]

Group Influence: The magnitude of a group's interests, even when scaled by state size, does not account fully for that group's ability to overcome the free-rider problems associated with political lobbying. These sorts of problems have been incorporated into empirical research on nonpolitical collective action, most notably in dealing with joint maximization in oligopolistic markets. Available data permit us to address this issue in the political context. To reflect the likelihood that a group can surmount free-rider difficulties, we introduce Herfindahl indices by state for surface coal producers, underground coal producers, coal consumers, and environmental organizations (*HSURF, HUNDER, HCONSUME*, and *HENVIROS*).[14] Data do not permit similar measures for *SPLITRIGHTS* and *UNRECLAIMED*.[15] Herfindahl indices are negatively related to group size and positively related to the concentration of interests within a group. Accordingly, *HSURF* and *HCONSUME* are expected to have negative effects on *ANTISTRIP*, while *HUNDER* and *HENVIROS* should have positive impacts.

C. *Senator Ideology*

The final variable to be introduced into the examination of *ANTISTRIP* voting is some (arguably impure) measure of senators' own ideologies. Following the lines of previous research, we rely on the independent (but not disinterested) "pro-environment" rating scale of the League of Conservation Voters (LCV). This rating scale is based on 27 not-surface-mining-related Senate votes taken in the 95th Congress and deemed to be ideologically revealing by the LCV.[16] The

[12] The groups are Sierra Club, National Audubon Society, Environmental Defense Fund, Friends of the Earth, National Wildlife Federation, and Wilderness Society (data provided by Resources for the Future, Inc.,). The correlation between *ENVIROS* and total hunting and fishing licenses in a state (U.S. Department of Commerce, *Statistical Abstract*, 1979) is 0.37. The correlation with state budgets for parks and recreation (*Statistical Abstract*, 1977) is −0.13; and is 0.003 with visits to state and federal forests and parklands (data from Charles Goeldner and Karen Dicke, 1981).

[13] Data for *SPLITRIGHTS* and *UNRECLAIMED* are from ICF, Inc. and the *Survey of Current Business*.

[14] The *Keystone Coal Industry Manual* provides state mine-by-mine data; *HCONSUME* is based on coal-fired electric generating capacity as reported in National Coal Association, *Steam Electric Plant Factors* (1978), *HENVIROS* is based on the six environmental groups mentioned in fn. 12 above.

[15] The omission does not appear to be consequential; the memberships of *SPLITRIGHTS* and *UNRECLAIMED* are relatively small and easily self-identified. Our earlier paper examined free-rider problems affecting these groups by introducing squared values of *SPLITRIGHTS* and *UNRECLAIMED*. Results are insignificant and do not alter any conclusions of interest.

[16] Votes are from LCV, *How Senators Voted on Critical Environmental Issues* (1978) and are listed in our earlier paper.

LCV notion of environmentalism conforms well with the aforementioned moralistic values of an undeveloped environment.[17] Analogous to *ANTISTRIP*, the frequency of pro-environmental votes is transformed according to (1) and is denoted *PROLCV*. Reflecting the LCV's own stance, *PROLCV* is expected to be positively related to *ANTISTRIP*. The extremes of *PROLCV* are occupied by senators with reputations as ideologues—for example, Kennedy (D-MA), Culver (D-IA), Zorinsky (R-NE), and Hatch (R-UT). Of course, this observation begs the question of the purity of ideology.

To the extent, if any, *PROLCV* reflects pure ideology, a move from a lower to a higher *PROLCV* represents a move from a less to a more intense demand for ideological support of an undeveloped environment. The variable *PROLCV* is built up from dichotomous, pro- or anti-environment choices by senators. Holding other things constant, including the opportunity cost of shirking, senators with relatively more intense demands for ideological environmentalism will choose the "pro" position more frequently and, hence, will have higher *PROLCV* values.

To the extent *PROLCV* reflects apparent ideology that is in fact proxying for constituents' interests in SMCRA, *PROLCV* should exhibit significant collinearity with the other factors that explain *ANTISTRIP* voting. Of course, as in any econometric analysis, there may be left-out variables. If these are correlated with the (apparent) ideologies of senators, the hypothesis that pure ideology matters in specific-issue politics may be inappropriately accepted. Consequently, a major part of the effort undertaken in Section III is aimed at uncovering such a correlation. At this stage, it can only be noted that each of the interests appearing in the record of SMCRA lobbying efforts and suggested by theory has been identified to the extent allowed by the data: all voters at least have "homes" in the selected variables and account is taken of the nature of political organization. To be sure, the groupings of voters according to their interests in coal strip mining most likely do not correspond to the groupings ("constituencies") that originally got a senator elected—as a result of the "bundling" discussed in Section I. But, the capture models of regulation do not suggest that the search for left-out variables should begin with these original groupings. Insofar as these models are specific-issue models, the search for left-out variables should be guided by analysis of SMCRA's impacts on consumers' and producers' surplus.

D. *Interstate Lobbying and Logrolling*

Two aspects of the legislative process not included in our empirical model are worth noting here. First, the conception of senators' voting choices embodied in our analysis portrays "captured" senators as casting their ballots based on the likely impact of SMCRA on their own states' constituents—with better organized constituents getting more attention. This obscures the ability of voters to apply pressure across state lines. Data inadequacies do not permit us to formally incorporate this phenomenon. Nevertheless, to the extent that cross-state lobbyers must appeal to within-state impacts in order to be effective, the problem recedes. Moreover, there is no obvious reason why SMCRA-*specific* out-of-state interests should cluster around *PROLCV* or any other measure of ideology based on a bundle of a senator's *non*-SMCRA votes. Still, this possibility becomes a central object of investigation below.

[17] Joseph Sax (1980) provides the archetypical statement of ideological environmentalism:

> The preservationist is not an elitist who wants to exclude others, notwithstanding popular opinion to the contrary; he is a moralist who wants to convert them. He is concerned about what other people do in the parks not because he is unaware of the diversity of taste in the society but because he views certain kinds of activity as calculated to undermine the attitudes he believes the parks can, and should encourage. [p. 14]
>
> ...Engagement with nature provides an opportunity for detachment from the submissiveness, conformity, and mass behavior that dog us in our daily lives; it offers a chance to express distinctiveness and to explore our deeper longings. At the same time, the setting—by exposing us to the awesomeness of the natural world in the context of "ethical" recreation—moderates the urge to prevail without destroying the vitality that gives rise to it: to face what is wild in us and yet not revert to savagery. [p. 42]

A second aspect of the legislative process not covered explicitly by our model is logrolling. This in part reflects the paucity of help provided by theory and data that would allow measurement of the extent and direction of any logrolling and coalition-forming on the specific issue of SMCRA—of the hundreds of issues to choose from, which issue(s) would a logrolling senator trade his SMCRA vote for? Note, however, there are no apparent a priori reasons why logrolling would make ideology appear any more or any less important relative to constituent interests in explaining SMCRA voting. That is, the willingness of a senator to trade away *either* his constituents' interests *or* his own ideology should be negatively related to the political strength of those interests and the intensity of his ideological preferences, respectively. Furthermore, in the absence of pure ideology, the hypothesis that ideology matters will incorrectly be accepted only if two conditions hold:

1) The non-SMCRA interests that are in fact served when a senator votes against his constituents' SMCRA-specific interests are systematically related to the interests that were served (either indirectly through logrolling or directly) by the senator's voting on the issues from which *PROLCV* is constructed.

2) At the same time, for the senator buying SMCRA votes by giving up his constituents' interests on other issues, SMCRA-specific constituent interests must be systematically related to the interests that are being served (either indirectly or directly) by his *PROLCV* voting.

With *PROLCV* issues ranging from the regulation of nitrogen oxide emissions from automobiles and the elimination of phosphates in dishwashing detergent to expansion of Redwood National Park and charging congressional staffers for parking privileges, satisfaction of these two requirements seems somewhat implausible. Nevertheless, we address this important question empirically in Section III—for the cases of both *PROLCV* and measures of ideology that are completely unrelated to the environment.

Finally, for the second condition above to hold without introducing collinearity between *PROLCV* and the included interest variables used to explain *ANTISTRIP*, the SMCRA-specific constituent interests being served by "buying" senators must be unrelated to the interests we have been able to identify. Again, a reading of the history of SMCRA provides little help in identifying such potent left-out variables. Still, this implication suggests a further object for empirical investigation.

E. *Initial Results*

It is clear that the task of isolating the determinants of legislative voting on an economic issue such as SMCRA is extraordinarily complex. As a first cut, we present the "standard" analysis that has been applied in previous research. Table 1 compares the Capture Model argued for by the economic theory of regulation (i.e., *PROLCV* is excluded) and a Capture-plus-Ideology Model that includes a variable (*PROLCV*) intended to account for senators' ideological preferences. Both models lend support to a multigroup (for example, see Peltzman, 1976) capture theory of politics—perhaps amended to include capture by ideologues. Noncoal beneficiaries of the environment, coal consumers, underground coal producers, and surface coal producers all appear to have appreciably influenced senators' voting on SMCRA; and interest groups' organizational capacities appear to have generally pushed senators in expected directions, although without especially strong statistical significance.

The most striking result of Table 1 is the sharp increase in explanatory power that results from the introduction of *PROLCV*. Indeed, it is this type of result that led Kau and Rubin, and Kalt (1981) to conclude that pure ideology was at work in legislative politics. The foregoing discussion, however, argues that such a conclusion is premature. Section III proceeds with a dissection that accounts for the extent to which an extended array of constituent characteristics (including *ENVIROS*) can explain *PROLCV*. At this stage, we can only note that, while the behavior of selected coefficients indicates that *PROLCV* is not completely orthogonal to

TABLE 1—THE DETERMINANTS OF ANTI-STRIP-MINING VOTING IN THE U.S. SENATE[a]

Explanatory Variable	Capture Model	Capture-plus-Ideology Model[b]
PROLCV	–	0.466
		(10.05)
		[0.65]
MC	−0.513	−0.375
	(−4.78)	(−3.47)
		[−0.22]
SURFRES	−16.765	−17.198
	(−1.66)	(−1.71)
		[−0.57]
UNDERRES	12.512	14.132
	(2.09)	(2.37)
		[0.73]
SPLITRIGHTS	−26.546	68.488
	(−0.55)	(1.40)
		[0.12]
ENVIROS	83.375	0.501
	(4.48)	(0.02)
		[0.00+]
UNRECLAIMED	0.019	0.015
	(3.77)	(3.03)
		[0.22]
CONSUME	−0.350	−0.440
	(−1.46)	(−1.83)
		[−0.13]
HSURF	−0.294	0.017
	(−1.24)	(0.07)
		[0.00+]
HUNDER	0.305	0.150
	(1.10)	(0.54)
		[0.03]
HENVIROS	1.935	−1.286
	(1.78)	(−1.14)
		[−0.07]
HCONSUME	−0.486	−0.261
	(−2.42)	(−1.29)
		[−0.08]
Constant	−0.154	1.414
	(−0.33)	(2.86)
$\bar{R}^2$	0.45	0.74
Condition-Stat.	25.99	27.47

[a] Dependent variable is *ANTISTRIP*; *t*-statistics are shown in parentheses.
[b] *Beta* coefficients are shown in brackets.

the set of other explanatory factors, the variable's statistical significance and the condition statistics (David Belsley et al., 1980) indicate there is insufficient collinearity to justify the conclusion that *PROLCV* is merely a proxy for the constituent interests identified by the capture theory.

III. Separating Interests and Ideology

Even if *PROLCV*'s sources—indiscernible constituent interests or the elusive notion of senatorial concerns for the public interest—are unclear, its explanatory power is striking. In the following analysis, we attempt to pry open the black box of ideology from a number of different angles. We first look for a purer measure of senators' own demands for altruistic, publicly interested behavior. Second, after isolating that portion which is most clearly pure, we examine the relative importance of the pure and remaining, arguably interest-proxy, parts of ideology. We then investigate the apparent interest-proxy part of ideology more closely. Finally, to assay whether ideological consumption is economically rational, we subject it to a revealing comparative statics test.

A. SMCRA *Voting and Social Issue Ideology*

Among the many issues senators vote on are certain moral and ethical matters around which economic-interest lobbying is infrequent. Examples include such issues as child pornography, the neutron bomb, and capital punishment. While Peltzman (1982) rejects ideology as an explanation for voting on economic issues such as SMCRA, he suggests that voting on noneconomic socio-ethical questions is especially likely to be based on individuals' preferences for moral rectitude, that is, pure ideology. Indeed, he finds evidence that voting on such issues reflects senators' own preferences more than does voting on "pocketbook" issues. Following this line of reasoning, we throw *PROLCV* out of the analysis of *ANTISTRIP* and replace it with measures of senators' social issue ideology. These measures are based on senators' voting on, for example, increased penalties for trafficking in child pornography, expanding the applicability of the death penalty, allowing the immigration of avowed communists, and "giving away" the Panama Canal. (Subsequent sections assess whether these measures actually reflect left-out interests and/or capture by ideological constituents.)

Two types of social issue ideology variables are examined. First, *PROLCV* is replaced by (dichotomous) voting on individual issues. Second, two indexes are created (according to (1)) from the sample of individual votes. The sample includes all votes taken in the 95th Congress that could be identified as general socio-ethical questions, uncontaminated by pocketbook concerns.[18] The issues thus identified are the column headings in Table 2. Selection was based on a priori judgment; that is, there was no econometric "fishing."

One of the indexes, the "SI (Social Issue) Index," is based on 34 votes dealing with the 12 non-Panama Canal issues indicated in Table 3. The second index, the "Panama Canal Index," is based on a sample of 25 procedural votes taken during the ratification process for President Carter's Panama Canal Treaty. The ceding of the Panama Canal was selected because a reading of the legislative history indicates that first, it was probably the most striking recent case in which conservatives "stonewalled it" against liberals; and second, no identifiable economic interest groups were coalesced by the issue.

The ideological content of the social issue votes cuts along modern liberal/conservative lines. To provide consistency to expected signs, senators are assigned a value of unity when voting the liberal position (as defined by, for example, the ADA) and a value of zero otherwise. It turns out that politicians consistently package liberalism and environmentalism together—the correlation between the LCV's and the ADA's rating scales is 0.94. Accordingly, if the apparent ideology embodied in *PROLCV* is, in fact, as pure as the ideology expressed in voting on socio-ethical matters, the social issue measures should have strongly positive effects on *ANTISTRIP*. Moreover, overall estimation results should closely resemble those found when using *PROLCV*.

Table 2 reports representative results when social issue ideology replaces *PROLCV* in the explanation of *ANTISTRIP*. Table 3 shows results of interest when an *individual* vote on one of the socio-ethical issues covered by the SI Index replaces *PROLCV*. The striking finding is how well voting on an issue with as much pocketbook content as SMCRA can be explained by senators' positions on the death penalty, sex education, the neutron bomb, the ceding of the Panama Canal, the immigration of avowed communists, and so on. In every case, the social issue variable has a strongly positive impact on *ANTISTRIP*.[19] Furthermore, the explanatory power of the Capture-plus-Ideology Model is remarkably similar when social issue voting replaces *PROLCV*. As might be expected, this is most evident when indexes are used. The thrust of these first results tilts toward the interpretation of *PROLCV* as reflecting relatively pure ideology.

B. *Isolating the Purest Part of Senator Ideology*

In an analysis of voting on the aggregate bundle of issues senators faced in the 96th Congress, Peltzman (1982) demonstrates that, at least on economic issues, measures of senatorial ideology (for example, ADA rating scales and senators' choices of party affiliation) stand in for a detailed list of constituent characteristics that plausibly correspond to their underlying economic interests. In the following analysis, we assume that this finding applies in the particular case of SMCRA. We further allow capture by ideological constituents. We then split measured ideology into that part that can be explained by constituent characteristics and the remaining senator-specific component. Our primary object is to examine whether the latter has any explanatory power.

We first estimate *PROLCV* and the SI Index as functions of the types of factors suggested by Kau and Rubin, and by Peltzman (1982)—factor such as general constituent characteristics and each senator's *PARTY* (Democrat = 1). Included constitu-

[18] Votes are from *Congressional Quarterly Almanac* (1977, 1978) and are described in our earlier paper.

[19] The *beta* coefficient for the SI Index (0.57) is the second largest in the model.

TABLE 2—IDEOLOGY, SOCIAL POLICIES, AND SMCRA VOTING: REPRESENTATIVE RESULTS[a]

Explanatory Variable	Communist Immigration	Death Penalty	SI Index	Panama Canal Index	*PROLCV*[b]
IDEOLOGY	0.842	1.013	0.296	0.193	0.466
	(5.92)	(7.19)	(9.62)	(8.93)	(10.05)
MC	−0.503	−0.464	−0.372	−0.434	−0.375
	(−4.22)	(−4.23)	(−3.44)	(−4.02)	(−3.47)
SURFRES	−21.157	−15.152	−16.204	−16.845	−17.198
	(−2.07)	(−1.46)	(−1.61)	(−1.67)	(−1.71)
UNDERRES	15.519	11.025	12.201	13.079	14.132
	(3.43)	(1.77)	(2.04)	(2.17)	(2.37)
SPLITRIGHTS	−1.077	−16.656	19.663	45.120	68.488
	(−0.02)	(−0.29)	(0.41)	(0.93)	(1.40)
ENVIROS	41.602	61.048	21.450	44.194	0.501
	(2.02)	(3.13)	(1.09)	(2.29)	(0.02)
UNRECLAIMED	0.014	0.022	0.011	0.013	0.015
	(2.75)	(3.86)	(2.10)	(2.46)	(3.03)
CONSUME	−0.223	−0.301	−0.280	−0.293	−0.440
	(−0.88)	(−1.14)	(−1.16)	(−1.18)	(−1.83)
HSURF	−0.048	−0.556	−0.380	−0.428	0.017
	(−0.19)	(−2.19)	(−1.60)	(−1.77)	(0.07)
HUNDER	0.159	0.435	0.416	0.407	0.150
	(0.53)	(1.43)	(1.50)	(1.39)	(0.54)
HENVIROS	0.164	1.370	0.192	0.962	−1.286
	(0.13)	(1.18)	(0.17)	(0.87)	(−1.14)
HCONSUME	−0.466	−0.138	−0.138	−0.197	−0.261
	(−2.21)	(−0.64)	(−0.67)	(−0.96)	(−1.29)
Constant	0.343	−0.614	0.730	0.111	1.414
	(0.69)	(−1.23)	(1.53)	(0.24)	(2.86)
$\bar{R}^2$	0.54	0.66	0.71	0.67	0.74

[a] See fn a, Table 1.

[b] From Table 1, Capture-plus-Ideology Model.

ent characteristics consist of demographic variables *and* measures intended to reflect constituents' independent ideological interests. By employing demographic variables, we are accepting the methodology of those who have found that senators' pure ideological goals play no role in legislative politics (Peltzman, 1982) and that demographic variables provide suitable proxies for constituents' underlying economic interests.[20]

The variables *ENVIROS* and *HENVIROS* are included to reflect constituents' (environmental) ideological interests. More generally, constituent ideology may be reflected by a measure such as the percentage of the state's vote going to McGovern (*MCGOV*) in the 1972 presidential election. The recognized hopelessness of McGovern's candidacy at election time probably made voting on the basis of investment motives unusually fruitless, and, by Stigler's (1972) argument, votes cast would uncommonly reflect ideological consumption motives.[21] Thus, *MCGOV* is

[20] Becker (1983) also argues that demographics provide natural groupings for constituents' interests. However, Arrow's 1963 General Possibility Theorem and the associated literature on coalitions, cycling, and electoral equilibria (see Dennis Mueller, 1979, for an excellent survey) suggest less confidence in the implicit econometric assumption (embodied in this work and elsewhere) that states' constituent characteristics can be mapped uniquely into their choices of elected representatives. Such concerns over uniqueness and existence, however, do not provide guidance as to why we find below that the explanatory power of constituent interests improves as a senator approaches his next electoral test (if that senator is not retiring!). This is consistent, however, with the interpretation of at least the part of the ideology measures which is *not* related to constituent characteristics as ideological shirking (see Section III, Part D, below).

[21] George McGovern was president of the ADA in 1976–78.

TABLE 3—IDEOLOGY, SOCIAL POLICIES, AND SMCRA VOTING: SOCIAL ISSUE VOTES[a]

Issue:	Coefficient	$\bar{R}^2$
Communist Immigration	0.842 (5.92)	0.54
Death Penalty	1.013 (7.19)	0.66
Pardon Draft Resisters	0.902 (6.53)	0.57
Sex Education	1.177 (7.70)	0.64
Neutron Bomb	0.893 (5.90)	0.54
School Desegregation	0.945 (6.67)	0.62
Abortion	0.591 (4.73)	0.53
Child Pornography	0.623 (3.52)	0.44
Pregnancy Disability	1.209 (6.07)	0.56
Pregnancy Discrimination	1.319 (7.11)	0.61
Cuba in Africa	0.882 (5.54)	0.62
Loans to Communists	0.888 (6.45)	0.55
Panama Canal Treaty	1.185 (8.64)	0.65

[a] Capture-plus-ideology specification, Table 2; t-statistics are shown in parentheses.

included in the explanation of *PROLCV* and the SI Index. Arguably, it is correlated with left-out constituent characteristics; but the exact source of *MCGOV* is irrelevant to our purpose of isolating that portion of *senators'* voting which is not related to some constituent (economic *or* ideological) interest. The source of *PARTY*, on the other hand, is a matter of concern.

If the list of included constituent characteristics is complete, *PARTY* reflects senator-specific "non-economic factors" (Peltzman, 1982) such as a senator's world view (liberal or conservative) at the time of the party affiliation choice. In the absence of a guarantee that our set of constituent characteristics is complete, we perform our analysis from both perspectives: party as ideology and party as proxy for unidentified constituent interests. The latter interpretation is rendered less plausible by the inclusion of *MCGOV*.

The estimated models of *PROLCV* and the SI Index (see our earlier paper) closely resemble the patterns of correlation reported by Kau and Rubin, Kalt (1981), and Peltzman (1982).[22] Their primary role here is to allow the breaking down of ideology. Specifically, *PROLCV* and the SI Index are split into that part predicted by constituent (economic and ideological) interests and a residual component. The fitted component (denoted the constituent part) of *PROLCV* and the SI Index is obtained both with *PARTY* included as a possible left-out interest proxy and with *PARTY* excluded. Correspondingly, the residual, senator-specific component consists of either each model's prediction errors alone or these errors plus the fitted *PARTY* effect. If there is any senator ideology at work in *PROLCV* or SI voting, this senator-specific component is the purest part that might be isolated.

The constituent part and the senator-specific component of *PROLCV* and, alternatively, the SI Index are entered as separate explanatory variables in the model of *ANTISTRIP*. Clearly, given previous results, the constituent part of the ideology measures should have a significant effect on *ANTISTRIP*. Of central interest, however, is the effect of the senator-specific part of the ideology measures. If senators are well policed on strip mining issues or if, when shirking, they do so nonideologically and independent of constituent interests, this variable should not be systematically related to *ANTISTRIP* voting. The alternative hypothesis that the senator-specific variable has a significantly positive impact on *ANTISTRIP* depends on two conditions. First, the variable must be isolating senators' pursuit of their own altruistic, publicly interested goals on general environmental issues (*PROLCV*) or socio-ethical matters (the SI Index).

[22] Variables included are *PARTY*, *MCGOV*, *ENVIROS*, per capita income, voter educational attainment, the fraction of state personal income generated by manufacturing, voter age, the urban-rural distribution of voters, the rate of growth of the state economy, and a southern dummy. Results of interest here are insensitive to lengthening the list of variables to include, for example, racial characteristics, unionization, and the blue-collar/white-collar split.

TABLE 4—LCV IDEOLOGY AND CONSTITUENT INTERESTS IN SMCRA VOTING[a]

	Excluding Party & Error Part of *PROLCV*		Party & Error Part of *PROLCV*		Excluding Error Part of *PROLCV*		Error Part of *PROLCV*	
Explanatory Variable	Coefficient	*Beta*	Coefficient	*Beta*	Coefficient	*Beta*	Coefficient	*Beta*
Senator-Specific Part of *PROLCV*								
Party & Error	–	–	0.442 (7.73)	0.45	–	–	–	–
Error Only	–	–	–	–	–	–	0.401 (4.89)	0.30
Constituent Part of *PROLCV*	0.615 (6.45)	0.48	0.527 (5.49)	0.41	0.570 (8.83)	0.56	0.511 (7.78)	0.50
MC	−0.422 (−3.91)	−0.24	−0.370 (−3.42)	−0.21	−0.483 (−4.50)	−0.28	−0.389 (−3.57)	−0.22
SURFRES	−25.826 (−2.54)	−0.86	−18.427 (−1.80)	−0.61	−24.446 (−2.42)	−0.81	−18.615 (−1.83)	−0.62
UNDERRES	18.634 (3.08)	0.96	14.894 (2.46)	0.77	17.467 (2.91)	0.90	14.860 (2.47)	0.77
SPLITRIGHTS	66.982 (1.33)	0.12	76.481 (1.52)	0.13	80.400 (1.62)	0.14	75.804 (1.53)	0.13
ENVIROS	−8.770 (−0.37)	−0.03	−7.932 (−0.34)	−0.03	6.150 (0.30)	0.02	−2.794 (−0.14)	−0.01
UNRECLAIMED	0.016 (3.15)	0.24	0.015 (2.97)	0.22	0.017 (3.49)	0.25	0.015 (3.06)	0.22
CONSUME	−0.493 (−2.04)	−0.15	−0.455 (−1.89)	−0.14	−0.544 (−2.26)	−0.16	−0.465 (−1.92)	−0.14
HSURF	0.009 (0.04)	0.00+	0.043 (0.18)	0.01	−0.115 (−0.48)	−0.03	0.009 (0.04)	0.00+
HUNDER	0.373 (1.35)	0.08	0.168 (0.60)	0.04	0.571 (2.05)	0.12	0.223 (0.78)	0.05
HENVIROS	−0.630 (−0.54)	−0.04	−1.474 (−1.27)	−0.08	−0.883 (−0.78)	−0.05	−1.379 (−1.21)	−0.08
HCONSUME	−0.139 (−0.67)	−0.04	−0.225 (−1.08)	−0.07	−0.032 (−0.16)	−0.01	−0.205 (−0.98)	−0.06
Constant	1.706 (3.10)	–	1.590 (2.89)	–	1.190 (2.42)	–	1.454 (2.93)	–
$\bar{R}^2$	0.57		0.74		0.67		0.74	

[a]See fn. a, Table 1.

Second, senators must have pursued these goals in their voting on the specific issue of coal strip mining policy.

Table 4 reports estimated *ANTISTRIP* models under the fitting and splitting of *PROLCV*. Table 5 reports the analogous case for SI ideology. As anticipated, the constituent parts of both *PROLCV* and the SI Index have a strongly positive influence on *ANTISTRIP*. Independent of the interpretation of *PARTY* (i.e., ideology or interest proxy), the senator-specific measure also has a highly significant and positive effect on *ANTISTRIP* voting. In fact, in every case, its inclusion appreciably improves the explanatory power of the model.[23] Further, it seems noteworthy that these conclusions are not obviously weaker when looking at the senator-specific part of *social issue* voting.

The results of Tables 4 and 5 may, of course, be due to the exclusion of some

[23] In every case, the hypothesis that the senator-specific variable has no effect on the model's explanatory power is rejected at above the 99 percent confidence level.

TABLE 5—SOCIAL ISSUE IDEOLOGY AND CONSTITUENT INTERESTS IN SMCRA VOTING[a]

Explanatory Variable	Excluding Party & Error Part of Index		Party & Error Part of Index		Excluding Error Part of Index		Error Part of Index	
	Coefficient	*Beta*	Coefficient	*Beta*	Coefficient	*Beta*	Coefficient	*Beta*
Senator-Specific Part of Index								
Party & Error	–	–	0.288	0.44	–	–	–	–
			(7.86)					
Error Only	–	–	–	–	–	–	0.233	0.29
							(4.93)	
Constituent Part of Index	0.452	0.38	0.326	0.27	0.501	0.52	0.393	0.41
	(5.55)		(3.94)		(8.44)		(6.20)	
MC	−0.432	−0.25	−0.369	−0.21	−0.473	−0.27	−0.389	−0.22
	(−4.00)		(−3.40)		(−4.41)		(−3.58)	
SURFRES	−20.969	−0.70	−16.575	−0.55	−20.761	−0.69	−17.593	−0.58
	(−2.08)		(−1.64)		(−2.06)		(−1.74)	
UNDERRES	15.148	0.78	12.433	0.64	14.765	0.76	12.984	0.67
	(2.53)		(2.07)		(2.47)		(2.17)	
SPLITRIGHTS	16.589	0.03	22.087	0.04	40.386	0.07	31.148	0.05
	(0.34)		(0.45)		(0.83)		(0.64)	
ENVIROS	18.950	0.06	17.641	0.06	19.517	0.06	14.273	0.05
	(0.86)		(0.80)		(0.97)		(0.71)	
UNRECLAIMED	0.016	0.24	0.011	0.16	0.016	0.24	0.012	0.18
	(3.10)		(2.08)		(3.23)		(2.27)	
CONSUME	−0.355	−0.11	−0.282	−0.09	−0.413	−0.12	−0.315	−0.10
	(−1.48)		(−1.17)		(−1.72)		(−1.31)	
HSURF	−0.186	−0.04	−0.369	−0.09	−0.270	−0.06	−0.354	−0.08
	(−0.78)		(−1.54)		(−1.14)		(−1.49)	
HUNDER	0.457	0.10	0.426	0.09	0.665	0.14	0.507	0.11
	(1.64)		(1.53)		(2.38)		(1.80)	
HENVIROS	−0.105	−0.01	0.066	0.01	−0.744	−0.04	−0.290	−0.02
	(−0.09)		(0.06)		(−0.66)		(−0.26)	
HCONSUME	−0.178	−0.05	−0.121	−0.04	−0.020	−0.01	−0.063	−0.02
	(−0.86)		(−0.58)		(−0.10)		(−0.30)	
Constant	1.227	–	0.823	–	1.022	–	0.916	
	(2.31)		(1.55)		(2.09)		(1.87)	
$\bar{R}^2$	0.54		0.71		0.65		0.72	

[a] See fn. a, Table 1.

as-of-yet-unidentified SMCRA-specific constituent interest—perhaps even an out-of-state interest, or a non-SMCRA interest served indirectly by logrolling. Tables 4 and 5, however, reduce the likelihood that such a variable exists. That is, if such a variable were to exist, it would have to be relatively orthogonal to not only states' identified SMCRA-specific interests, but also to the constituent interest variables used in the explanation of *PROLCV* and the SI Index. Moreover, it would have to be *simultaneously* and causally (in the capture sense) related to voting on coal strip-mining regulation, general environmental issues, sex education, the neutron bomb, child pornography, etc. Pending further search for such a variable, Tables 4 and 5 further suggest that ideology probably should not be excluded from analyses of particular-issue politics.

C. *The Role of Left-Out Constituent Characteristics*

The preceding analysis has assumed that a large part of senators' apparent ideology is correlated with causal constituent (economic or ideological) interests left out of the basic capture-plus-ideology explanation of *ANTISTRIP*. As noted in Section I, however, such correlations could arise for two very different reasons. On the one hand, it may be happenstance: what looks like senator ideology is the reflection of left-out constituent economic or ideological interests that actually have direct effects on SMCRA voting. On the other hand, *PROLCV* and the SI Index may be entirely pure senator ideology, unrelated to the interests constituents have in the *specific* issue of SMCRA, but related to broad constituent characteristics via Senate elections and voters' interests in putting the "right" ideologies into office—a la Downs.

The problem of inferring causation from correlation is notably difficult. In the case at hand, we approach this problem by focusing on whether there are different implications: 1) when the variables that are correlated with measures of ideology are operating directly on *ANTISTRIP*, as opposed to 2) when policing is imperfect and they operate indirectly through their impact on the types of senators elected (and, hence, the ideologies manifested by shirking senators).

One such difference in implications concerns the way the constituent characteristics (i.e., demographics, ideology, and perhaps *PARTY*) that explain *PROLCV* and the SI Index enter the explanation of *ANTISTRIP*. The same set of characteristics may play a causal role in all three types of voting. The obvious differences in the economic content of SMCRA, general environmental issues, and socio-ethical matters, however, suggest that the roles played by the individual variables in this set will not be identical across the three cases in terms of their size, signs, and significance. Consequently, if the variables that make up the constituent part of ideology play direct causal roles in SMCRA voting, their explanatory power will be higher if they are not constrained to enter the *ANTISTRIP* model in the same linear combinations they have in the explanation of *PROLCV* or the SI Index. On the other hand, the constituent part of ideology may in fact be pure senator ideology in the particular case of SMCRA *and* the route of influence by the set of general constituent characteristics may thus be the indirect Downsian route. If so, constraining these characteristics to enter the *ANTISTRIP* model in the same way they enter the explanation of the ideology measures should not affect their explanatory power.

Results when the general constituent characteristics are constrained to enter the *ANTISTRIP* model as the fitted constituent parts of *PROLCV* and the SI Index are what is reported in Tables 4 and 5. Table 6 compares the explanatory power of the constrained characteristics with that obtained in the unconstrained case (see our earlier paper for full results). Table 6 indicates that, of course, the R^2 is higher when the general constituent characteristics are not con-

TABLE 6—IDEOLOGY AND THE PATH OF GENERAL CONSTITUENT CHARACTERISTICS IN ANTISTRIP VOTING

	LCV Ideology: Constituent Characteristics		SI Ideology: Constituent Characteristics	
Model	Unconstrained	Constrained[a]	Unconstrained	Constrained[b]
Party as Ideology:				
R^2	0.78	0.77	0.79	0.75
$\bar{R}^2$	0.73	0.74	0.73	0.71
F-Test[c]	84%		14%	
Party as Interest Proxy:				
R^2	0.79	0.77	0.79	0.76
$\bar{R}^2$	0.73	0.74	0.74	0.72
F-Test[c]	89%		21%	

[a] From Table 4.

[b] From Table 5.

[c] The level of significance at which equality of the two regressions may be rejected. Lower values signify increased confidence in rejection of equality.

strained. In the SI Index case, the adjusted R^2 also improves when the characteristics are unconstrained—although only slightly. In the *PROLCV* case, however, the adjusted R^2 actually falls when constituent characteristics are unconstrained. In each case, classical tests of the equality of the constrained and unconstrained models provide little confidence that equality can be rejected.

The general constituent characteristics that might have been supposed to be directly affecting SMCRA voting appear to be operating indirectly through the pure ideology of elected and shirking senators. That is, *PROLCV* and the SI Index do not appear to be standing in for otherwise unidentified economic or ideological interests that might have captured senators in their voting on the particular issue of SMCRA. The suggested conclusion is that a senator's overall bundle of votes reflects the constituent preferences that were expressed on election day. The interests of those same constituents, when those interests are subsequently reshuffled by the prospective effects of a particular decision on the Senate floor, however, do not fully control the senator on that decision. The capture theory may work well as a theory of elections, but not so well as a theory of issue-specific politics.

D. *Electoral Policing and Ideological Shirking*

If the evident roles of *PROLCV* and the SI Index in SMCRA voting largely reflect senators' own ideological preferences, they do so because of imperfect issue-specific policing. Accordingly, the extent to which shirking senators' notions of the public interest enter their voting behavior should be systematically related to the opportunity costs of shirking. The investigation of this cost is a promising source of comparative statics tests for the presence, importance, and rationality of ideology in representative politics. In general terms, the opportunity cost of shirking should vary directly with constituents' incentives to police (i.e., the producers' and consumers' surplus stakes); and vary inversely with organization-information costs and institutional impediments to monitoring. The cost of shirking should also be inversely related to the security, somehow measured, with which a senator intending to remain in office holds his or her seat. This security is plausibly related to such factors as the senator's margin of last victory, tenure, committee power (a major source of facilitation services), and personal wealth. These relationships obviously involve substantial simultaneity—tenure, for example, affords senators security, but they get tenure by serving their constituents.

The development of a full simultaneous model of shirking is considerably beyond the scope of this study. Nevertheless, one component of the cost of shirking that is clearly determined exogenously can be examined—the proximity of a senator's next electoral test (Robert Barro, 1973). In a world in which voters have positive discount rates and/or memories that decay, a senator can, in a sense, run down his or her security capital in midterm and build it back up again as the next election approaches. Ideological shirking should, therefore, be directly related to the time remaining in a senator's term.

Table 7 reports one test of this hypothesis. In our sample, 33 senators were up for reelection in 1978. If present, ideological shirking should have been less prevalent (*ceteris paribus*) among these senators than in the sample as a whole during the strip-mining voting of 1977. To examine this, we look for errors in the predictions of the Capture-plus-Ideology Model of *ANTISTRIP*. On any particular vote, the expected outcome, yea or nay, is given by the sign of the senator's predicted *ANTISTRIP* value. The absolute magnitude of this value suggests the confidence to be placed in the expected outcome. An error is inferred when an actual outcome on a vote differs from the expected outcome. An error is interpreted as running against a senator's ideology when its sign is opposite to the sign of the senator's LCV rating.

Twelve errors are made in predicting SMCRA voting in 1977. Of these, approximately 50 percent were made by the 33 percent of the senators who were up for reelection. Moreover, of the 6 up-for-reelection

TABLE 7—THE EFFECT OF ELECTORAL POLICING ON IDEOLOGICAL VOTING

Senator's Electoral Status	Incorrect Predictions in 1977[a]		of which:	Voted Against LCV Ideology	
	Number	Percent of Total		Number	Percent of Col. 1
Up for Reelection	6	50		5	83.3
Not Up for Reelection	6	50		3	50
Total	12			8	

[a]Predictions based on Table 1, Capture-plus-Ideology Model. "Incorrect" signifies sign of predicted *ANTISTRIP* differed from sign of actual *ANTISTRIP* voting in 1977.

senators predicted incorrectly, 5 were voting against their ideologies; and the remaining senator voting with his ideology was retiring. Of the 6 not-up-for-reelection errors, only 3 were represented by senators voting against their ideologies. The probability that this pattern was generated by chance is very low.[24] Thus, it appears that the proximity of the next election inhibits ideological shirking: senators shirk less as the policemen approach.

IV. Summary and Conclusions

The evidence thus far suggests the need for some broadening in the economic theory of politics. This theory has effectively precluded from its list of determining factors anything other than the parochial, narrowly self-interested objectives of policymakers' constituents. This reflects the fact that economic, capture models of regulation are largely institution-free. Our analysis has attempted to bring certain aspects of the property rights theory of institutions—policing costs, opportunism, appropriability—to bear on the legislative process. Specifically, we have asked whether slack in control of legislators is empirically important enough to warrant incorporation into positive models of politics. At the level of specific-issue policymaking, the answer appears to be yes. Our results also suggest that the discretionary consumption afforded policymakers by institutional slack is taken, to a substantial degree, in the form of rational altruistic-ideological promotion of self-defined notions of the public interest. At least in the case of federal coal strip-mining policy, this ideological shirking appears to have significantly affected the course of public policy.

If the concept of ideological shirking does prove to be significant, its usefulness will depend on the development of models that can predict the conditions (for example, types of issues, institutional settings, economic contexts) under which ideological shirking is likely to be an important phenomenon. Of course, it still may be that the phenomenon does not even exist. There may yet be constituent interests missing from this and previous analyses that will explain away ideology's importance in specific-issue politics. The search for these interests should continue. For now, it appears that the economic theory of regulation will have to keep the door open to ideological behavior.

REFERENCES

Abrams, Burton A. and Settle, Russell F., "The Economic Theory of Regulation and Public Financing of Presidential Elections," *Journal of Political Economy*, April 1978, *86*, 245–57.

Alchian, Armen and Kessel, Reuben, "Competition, Monopoly, and the Pursuit of Pecuniary Gain," in H. G. Lewis, ed., *Aspects of Labor Economics*, Princeton:

[24]Bearing in mind the small sample, a *Chi*-squared test indicates that the hypothesis that the contingencies in Table 7 were generated by chance can be rejected at above the 99 percent confidence level.

National Bureau of Economic Research, 1962, 156–83.

______ **and Demsetz, Harold,** "Production, Information Costs, and Economic Organization," *American Economic Review*, December 1972, *62*, 777–95.

Arrow, Kenneth J., "Gifts and Exchanges," *Philosophy and Public Affairs*, Summer 1972, *1*, 343–62.

Ashenfelter, Orley and Kelley, Stanley, Jr., "Determinants of Participation in Presidential Elections," *Journal of Law and Economics*, December 1975, *18*, 162–70.

Barro, Robert J., "The Control of Politicians: An Economic Model," *Public Choice*, Spring 1973, *14*, 19–42.

Barzel, Yoram and Silberberg, Eugene, "Is the Act of Voting Rational?," *Public Choice*, Fall 1973, *16*, 51–58.

Becker, Gary S., *The Economics of Discrimination*, Chicago: University of Chicago Press, 1957.

______, "A Theory of Social Interactions," *Journal of Political Economy*, December 1974, *82*, 1063–93.

______, "Altruism, Egoism, and Genetic Fitness," *Journal of Economic Literature*, September 1976, *14*, 817–26.

______, "Competition Among Pressure Groups for Political Influence," *Quarterly Journal of Economics*, August 1983, *98*, 371–98.

Belsley, David A., Kuh, Edwin and Welsch, Roy E., *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: Wiley & Sons, 1980.

Brennan, Geoffrey and Buchanan, James M., "Voter Choice and the Evaluation of Political Alternatives: A Critique of Public Choice," unpublished, Center for the Study of Public Choice, Virginia Polytechnic Institute and State University, 1982.

Breton, Albert, *The Economic Theory of Representative Government*, Chicago: Aldine Publishing, 1974.

Buchanan, James M. and Tullock, Gordon, *The Calculus of Consent*, Ann Arbor: University of Michigan Press, 1965.

De Alessi, Louis, "On the Nature and Consequences of Private and Public Enterprises," *Minnesota Law Review*, October 1982, *67*, 191–209.

Downs, Anthony, *An Economic Theory of Democracy*, New York: Harper and Row, 1957.

Ferejohn, John A., "On the Decline of Competition in Congressional Elections," *American Political Science Review*, March 1977, *71*, 166–76.

Fiorina, Morris and Noll, Roger G., "Voters, Legislators, and Bureaucracy: Institutional Design in the Public Sector," *American Economic Review Proceedings*, May 1978, *68*, 256–60.

Frey, Bruno S. and Lau, Lawrence J., "Towards a Mathematical Model of Government Behavior," *Zeitschrift fur Nationalokonomie*, December 1968, *28*, 355–80.

Gart, John J. and Zweifel, James, "On the Bias of Various Estimators of the Logit and its Variance with Application to Quantal Bioassay," *Biometrika*, June 1967, *54*, 181–87.

Goeldner, Charles R. and Dicke, Karen P., *Travel Trends in the United States and Canada*, Boulder: GSBA, University of Colorado, 1981.

Hirshleifer, Jack, "Competition, Cooperation and Conflict in Economics and Biology," *American Economic Review Proceedings*, May 1978, *68*, 238–43.

Jensen, Michael C. and Meckling, William H., "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure," *Journal of Financial Economics*, October 1976, *3*, 305–60.

Kalt, Joseph P., *The Economics and Politics of Oil Price Regulation*, Cambridge: MIT Press, 1981.

______, "The Costs and Benefits of Federal Regulation of Coal Strip Mining," *Natural Resources Journal*, October 1983, *23*, 893–915.

______ **and Zupan, Mark A.,** "Further Evidence on Capture and Ideology in the Economic Theory of Politics," unpublished, Harvard Institute of Economic Research, April 1983.

Kau, James B. and Rubin, Paul H., "Self-Interest, Ideology and Logrolling in Congressional Voting," *Journal of Law and Economics*, October 1979, *22*, 365–84.

Klein, Benjamin, Crawford, Robert G. and Alchian, Armen, "Vertical Integration, Appropriable Rents, and the Competitive

Contracting Process," *Journal of Law and Economics*, October 1978, *21*, 297–326.

Manne, Henry G., "Mergers and the Market for Corporate Control," *Journal of Political Economy*, April 1965, *73*, 110–20.

Mitchell, Edward J., "The Basis of Congressional Energy Policy," *Texas Law Review*, March 1979, *57*, 591–613.

Mueller, Dennis C., *Public Choice*, Cambridge: Cambridge University Press, 1979.

Olson, Mancur, *The Logic of Collective Action*, New York: Shocken Books, 1971.

Peltzman, Sam, "Toward a More General Theory of Regulation," *Journal of Law and Economics*, August 1976, *19*, 211–40.

______, "Constituent Interest and Congressional Voting," unpublished, University of Chicago Economic and Legal Organization Workshop, February 1982.

Phelps, Charles E., Book Review: "Kalt's *The Economics and Politics of Oil Price Regulation*...," *Bell Journal of Economics*, Spring 1982, *13*, 289–95.

Riker, William H. and Ordeshook, Peter C., "A Theory of the Calculus of Voting," *American Political Science Review*, March 1968, *62*, 25–42.

Rothenberg, Jerome, "A Model of Economic and Political Decision Making," in J. Margolis, ed., *The Public Economy of Urban Communities*, Washington: Resources for the Future, Inc., 1965.

Samuelson, Paul, "The Pure Theory of Public Expenditure," *Review of Economics and Statistics*, November 1954, *36*, 387–89.

Sax, Joseph L., *Mountains Without Handrails, Reflections on the National Parks*, Ann Arbor: University of Michigan Press, 1980.

Schulze, William D. et al., *The Benefits of Preserving Visibility in the National Parklands of the Southwest*, Washington: U.S. Environmental Protection Agency, 1981.

Stigler, George J., "The Economic Theory of Regulation," *Bell Journal of Economics*, Spring 1971, *2*, 3–21.

______, "Economic Competition and Political Competition," *Public Choice*, Fall 1972, *13*, 91–106.

Tiebout, Charles M., "A Pure Theory of Local Expenditures," *Journal of Political Economy*, October 1956, *64*, 416–24.

Tollison, Robert D. and Willett, Thomas D., "Some Simple Economics of Voting and Not Voting," *Public Choice*, Winter 1975, *24*, 43–49.

Tullock, Gordon, "A (Partial) Rehabilitation of the Public Interest Theory," unpublished, Center for the Study of Public Choice, Virginia Polytechnic Institute and State University, 1982.

Williamson, Oliver E., *Markets and Hierarchies: Analysis and Antitrust Implications*, New York: Free Press, 1975.

Zellner, Arnold and Lee, Tong H., "Joint Estimation of Relationships Involving Discrete Random Variables," *Econometrica*, April 1965, *33*, 382–94.

Congressional Quarterly, Inc., *Congressional Quarterly Almanac*, Washington, various years.

ICF, Inc., *Final Report: Energy and Economic Impacts of H.R. 13950 (Surface Mining Control and Reclamation Act of 1976)*, Washington, 1977.

League of Conservation Voters, *How Senators Voted on Critical Environmental Issues*, Washington, 1978.

McGraw-Hill, Inc., *Keystone Coal Industry Manual*, New York, various years.

National Coal Association, *Coal Data 1978*, Washington, 1979.

______, *Steam Electric Plant Factors*, Washington, 1978.

U.S. Department of Commerce, *Statistical Abstract of the United States*, Washington: USGPO, various years.

______, *Survey of Current Business*, Washington: USGPO, various months.

The U.S. Productivity Slowdown: A Case of Statistical Myopia

By MICHAEL R. DARBY*

The decline in American productivity has come to rival inflation as a major economic issue for public policy. Indeed something akin to panic has followed reports that labor productivity growth declined from an average annual rate of 2.6 percent over 1948–65 to 1.9 percent over 1965–73, and to 0.5 percent from 1973 to 1979.[1] This paper shows that the productivity panic is based upon statistical myopia, and that a careful analysis within the perspective of the entire twentieth century discloses no substantial variation in what is variously described as growth in total factor productivity or technical progress.

The argument is made in two parts. First, three major subperiods are identified: 1900–29, 1929–65, and 1965–79. It is noted first that the early and late periods are very similar to each other and are characterized, in comparison with the middle period, by rapid growth in labor force and a less than equal increase in growth in real output so that measured labor productivity falls. This picture changes dramatically when allowances are made for age, sex, education, and immigration to obtain a quality-adjusted labor force. The differential between the growth rates of output and quality-adjusted hours worked is essentially identical across the three periods. Thus simple demographic adjustments eliminate any secular decline in technical progress.

The second part of the argument focuses on variations in productivity growth trends within the middle and later periods. It is shown that very slow growth in the capital-labor ratio from 1929 to 1948 accounts for very slow labor productivity growth during that subperiod as well as very rapid growth during 1948–65. That is, the very rapid growth of 1948–65 resulted from our poverty in 1948 and did not reflect desirable economic conditions. The more rapid productivity growth in 1965–73 as compared to 1973–79 is explicable by measurement error due to underreporting of price increases and hence overreporting of output increases due to President Nixon's price control program. It is of particular interest that the oil increases of 1973–74 do not appear to have played a major role in slowing productivity growth.

The analysis was made possible by the development of a historical data base on productivity, labor force, and employment reported elsewhere.[2] Unfortunately, complete census and immigration data required for the demographic adjustments were only available through 1979 when the data base was compiled in 1982. This is not a significant problem for analysis of longer-run trends, since 1979 is the last year available as of writing (1983) which is characterized by an approximately normal unemployment

*Department of Economics, University of California, Los Angeles, CA 90024. This research was begun with financial support received from the UCLA Institute of Industrial Relations and the Foundation for Research in Economics and Education. It has been completed as part of the National Bureau of Economic Research's Project on Productivity and Industrial Change in the World Economy, and with the support of the National Science Foundation (grant SES-8207336). An earlier version was presented as the Goldwater Lecture at the Arizona State University, September 24, 1982. Francine Blau, Barry Chiswick, Charles Cox, Roger Craine, Sebastian Edwards, Dan Friedman, John Haltiwanger, Ed Leamer, Axel Leijonhufvud, James Pierce, Mark Plant, Ken Sokoloff, and Ezio Tarantelli provided valuable comments. Charles S. Morris, Elizabeth M. Landau, Frances R. Hammond, Marjorie Rose, and Zaki Eusufzai provided able research assistance. Any opinions expressed are my own and not necessarily those of the funding organizations. This is not an official publication of the NBER.

[1] Productivity trend growth rates are normally computed from high-employment to high-employment year to avoid the large cyclical variations in productivity analyzed by such authors as Walter Oi (1962), Ray Fair (1969), Robert Solow (1973), Christopher Sims (1974), Robert J. Gordon (1979), and Charles Morris (1983). The figures in the text are for the private-nonfarm-output-per-hour-paid-for definition of labor productivity and are computed from data in the U.S. Council of Economic Advisers (1980, pp. 246).

[2] See my 1984 study.

TABLE 1—AVERAGE ANNUAL GROWTH RATES OF PRIVATE EMPLOYMENT, PRIVATE HOURS, GROSS PRIVATE PRODUCT, AND PRODUCTIVITY, 1900–79

Period	Private Employment (*PE*)	Average Hours Worked (*AHWP*)	Private Hours Worked (*THWP*)	Gross Private Product (*GPP*)	Hourly Productivity (*GPP/THWP*)	Employee Productivity (*GPP/PE*)
1900–79	1.40	−0.23	1.18	3.23	2.05	1.82
Major Periods:						
1900–29	1.77	−0.22	1.54	3.42	1.88	1.65
1929–65	0.87	−0.27	0.60	2.98	2.38	2.10
1965–79	2.02	−0.12	1.91	3.48	1.57	1.46
Subperiods:						
1900–16	2.09	0.05	2.14	3.64	1.50	1.56
1916–29	1.38	−0.56	0.81	3.15	2.34	1.78
1929–48	0.88	−0.17	0.71	2.28	1.57	1.40
1948–65	0.86	−0.40	0.47	3.75	3.28	2.89
1965–73	1.84	−0.21	1.63	3.90	2.27	2.06
1973–79	2.27	0.00	2.28	2.92	0.64	0.64

Source: See Data Appendix, Table A1.

Note: Units: Continuously compounded rates in percent per annum computed from the first to the last year of the period.

rate. Where cyclical productivity growth is considered in Section II, it was possible to use data through the first quarter of 1983. Extracts of the relevant data are contained in the Data Appendix.

I. Analysis of Longer-Period Trends

The broad trends of the twentieth century are summaiized in Table 1 for private employment (*PE*), average and total private hours worked (*AHWP* and *THWP*, respectively), gross private product (*GPP*), and private hourly and employee productivity (*GPP/THWP* and *GPP/PE*, respectively).[3] My 1984 study notes the periods 1900–29 and 1965–79 were characterized by rapid employment growth with immigration the relatively dominant factor in the early period, and the baby-boom new entrants relatively more important in the recent period. The intermediate period 1929–65 was marked by both tight limitation on immigration and a low rate of natural increase.[4] The three major periods were thus differentiated both by changes in immigration laws and by the postwar baby-boom generation's coming of age. Since the rate of decline in average hours worked was nearly constant, the changes in employment growth were the dominant factor determining variation in growth in total hours.

Each of the major periods has been divided roughly in half for later analysis.[5] At least for 1929–65 and 1965–79, the growth rates of private employment and total private hours are approximately the same in each subperiod as the mean for the respective

[3] As discussed below, the results are not very sensitive to the particular measures chosen. The ratio of *GPP* to total private hours worked indicates the largest 1965–79 private hourly productivity slowdown (0.65 percentage points) and is accordingly used. At the other extreme, total private hours paid would indicate only a 0.53 percentage point deceleration.

[4] Most of the growth in employment over 1900–29 appears to be concentrated in the years of massive immigration 1900–16, but the early data (especially for average hours worked) are not of sufficiently high quality for close analysis of movement within the 1900–29 period. Immigration was sharply limited by the "national origin" quota system which became fully operative in 1929 and was finally abolished by the Act of October 3, 1965. The average ratio of the annual flow of immigrants (age 16 and over) to civilian labor force was 1.42, 0.20, and 0.34 percent for 1900–29, 1930–65, and 1966–78, respectively.

[5] The break years 1948 and 1973 are the standard ones in the literature. The year 1916 was chosen as a convenient high-employment year.

major period. Thus each of these periods is roughly homogeneous in terms of labor developments. The unevenness of immigration and estimated average-hours-worked growth in the first 29 years of the century makes the period 1900–29 appear rather less homogeneous, but demographic adjustments discussed below eliminate most of the differences between the subperiods.

Focusing on the major periods in Table 1, note that 1965–79 is rather similar to 1900–29, not only in employment and hours growth but also in output and productivity growth. Compared to 1929–65, total hours growth is 1.0 to 1.3 percent higher in the earlier and later periods, while *GPP* growth is only 0.4 to 0.5 percent higher. Thus hourly productivity growth is recorded as 0.5 to 0.8 percent lower in 1900–29 and 1965–78 as compared to 1929–65. Correspondingly, growth in private output per person employed is 0.4 to 0.6 lower in the extreme periods than in the middle period. A possible explanation of the more rapid productivity growth in the middle period is that the hours and employment growth-rate declines in the middle period may be overstated due to failure to adjust for demographic changes, especially immigration and the baby boom.[6]

Measures of private hours do not adjust for differences in human capital, although the idea that an hour is an hour is as fallacious as the idea that a 1964 dollar equals a 1984 dollar. Although elaborate adjustments such as Peter Chinloy (1980) are precluded by limitations in the historical data, it is possible to make approximate adjustments for observed differences in productivity due to age, sex, education, and immigrant status.

Let us first consider the adjustment for the age-sex composition of the labor force. The private labor force estimates are divided into individual cells by sex and "young" (Y, under 25) or "old" (O, over 24). Standardizing on males over 24 and taking young males, young females, and old females as less productive because of differences in human capital, we compute age-sex adjusted private employment as

$$(1)\qquad APE = PE_{MO} + \alpha_1 PE_{MY} + \alpha_2 PE_{FY} + \alpha_3 PE_{FO},$$

where the subscripts M and F indicate sex. The αs are chosen to reflect differences in average hourly earnings. Using data in Edward Denison (1979, p. 33), this suggests α_1 of 0.53 to 0.50, α_2 of 0.43 to 0.41, and α_3 of 0.56 to 0.57.[7] I use 0.515, 0.42, and 0.565 as α_1, α_2, and α_3, respectively.

New immigrants to the United States on average earn substantially less than native-born Americans.[8] In part this reflects permanent differences in human capital endowments, but much of the difference is eliminated over time as the immigrants become acculturated. Since immigration is most important in the earlier period, Francine Blau's (1980) estimates based on 1909 data are used to adjust the foreign-born to native-born equivalents by the following formulas:[9]

$$(2)\qquad PE_{Ma} = PE_{MNa} + (1.01076)^{Z_M} \times (0.753) PE_{MIa},$$

$$(3)\qquad PE_{Fa} = PE_{FNa} + (1.01177)^{Z_F} \times (0.891) PE_{FIa},$$

[6] In earlier versions of this paper, an alternative capital-deepening hypothesis was considered. Along the lines of the neoclassical growth model (see, for example, Solow, 1970, pp. 17–38, or my 1979 book, pp. 105–15, 139–40, 440–41), the slower growth in labor should increase steady-state equilibrium capital-labor, capital-output, and output-labor ratios. However, we shall see in Section II that the capital-output ratio apparently was lower or the same in 1965 as compared to 1929. The apparent irrelevance of capital-deepening may reflect the operation of an efficient world capital market.

[7] The estimates are computed from relative earnings for 1929–70 and 1970–76, respectively, for finer age groups. The finer age groups were weighted for percentage of total hours worked. Note that the relative wages reflect not only pure age and sex differences in human capital but also the *relative* amounts of education, all taken as approximately constant. Changes in the average *level* of education are accounted for in equation (5) below.

[8] See Barry Chiswick (1978, 1979) and Francine Blau (1980). Previously Milton Friedman (1974) had noted that the rapid immigration of the early twentieth century reduced measured growth in real per capita income or, as here, labor productivity.

[9] These parameters were derived as follows. Blau (1980, p. 32) indicates gross log differentials in wages at

where PE_{sIa} and PE_{sNa} are the (unadjusted) private employment of foreign-born (I) and native-born (N) individuals of sex s and age-group a, and where Z_s is the average years since entry of foreign-born workers of sex s.[10] That is, a just-arrived male immigrant is assumed to equal 75.3 percent of a native-born male of the same age group, but this factor grows at a compound rate 1.076 percent per year spent in the country. The average years since entry was estimated according to the recursive formula

$$(4) \qquad Z_s = (0.5)(I_s/P_s) + (Z_{s,-1} + k_s)(1 - I_s/P_s),$$

where I_s is the inflow of (sex s) immigrants over the preceding year, P_s is the corresponding foreign-born population, and k_s is a number between 0 and 1 to allow for disproportionate frequency of emigration, death, and retirement among the less-recent foreign born. This formula says that the (midyear) average years since entry is a weighted average—weights I_s/P_s and $1-(I_s/P_s)$—of 0.5 year for those arriving in the last twelve months and $Z_{s,-1}+k_s$ for those previously arrived and remaining in the labor force. Benchmarks for 1909 and 1970 were computed from the Immigration Commission data for 1909 and from 1970 census data.[11] These benchmarks implied $k_M \approx k_F \approx 0.603$. The estimated values of Z_s are reported in the Data Appendix. Since the adjustments for years since migration of the foreign born were made before the adjustment for age and sex described in equation (1), the variable APE is in fact adjusted for all these factors.

The final demographic factor believed to be important in determining the human capital content of the labor force is education. It is assumed that human capital is increased by 7 percent per year of education,[12] so that quality-adjusted private employment is

$$(5) \qquad QAPE = (1.07)^E APE.$$

The education variable E is measured by the median school years completed by those 25 years and over.[13]

entry for Ethnic Group 2 (the relevant one for post-1900) of -0.351 for men and -0.183 for women. Now $e^{-0.351} = 0.704$ and $e^{-0.183} = 0.833$. These gross differentials were in part due to education, which is adjusted for separately below. Allowance is made for 1 year of education (see Chiswick, 1978, p. 907) or 7 percent which approximately squares with Blau's results for percent literate. (No allowance was made for differences in mean ages as preliminary calculations suggested any effect was negligible.) Thus the gross wage differential on entry used was $(1.07)(0.704) = 0.753$ for males and $(1.07)(0.833) = 0.891$ for females. The estimated effect on the log of real wages of years since migration was 0.0107 for males and 0.0117 for females, and $e^{0.0107} = 1.01076$ and $e^{0.0117} = 1.01177$. Using 1970 data for men, Chiswick (1978) obtained values implying an adjustment factor of about $(1.015)^{Z_M}(0.721)$ for males. A sensitivity check showed that substituting this factor made no significant difference to the results. In the early 1900's, immigration were primarily from Blau's Ethnic Group 2 (Irish, French Canadians, Southern and Eastern Europeans) while recent immigrants have been primarily of Latin American or Asian origin. While these groups are not strictly comparable, the foreign-born adjustments are primarily important in the earlier period.

[10] Unfortunately, no data were found to make differential adjustments by age group so the same factor was used for both young and old.

[11] The 1909 data came from U.S. Immigration Commission (1911, Table 56, pp. 1521 and 1528). The cells were assigned their mean values (0.5, 1.5, 2.5, 3.5, 4.5, 7.5, 12.5, and 17.5) except for the open cell (20 years and over) for which Blau's value of 30 years was used. The 1970 data were from U.S. Bureau of the Census (1973, Table 18, p. 466) using cell values of 2.65, 7.8, 12.8, 17.8, 22.8, 30.3, 40.3, and 53.1 years for the open (before 1925) cell. The last value was estimated by taking weighted averages of young immigrants arriving 1904 through 1924 who would be 65 or under in 1970.

[12] This value was taken from Chiswick (1978, p. 908) and appears reasonable in terms of such recent cross-section results as reported by James Smith and Finis Welch (1977). Yet lower rates of return to education would lower productivity growth in the early period relative to the two later periods, but would not have much effect on the main conclusion of this section: the absence of a secular decline in productivity growth.

[13] Mean years of education were not available, but if the difference is constant the substitution of the median will not affect the growth rates estimated below. John Folger and Charles Nam (1964) retroject the 1940 census data back to obtain median estimates for 1930, 1920, 1910. These values of 8.6, 8.4, 8.2, and 8.1 were extrapolated to 8.0 in 1900. Sources for more recent years are given in the Data Appendix. Log-linear interpolation was used to fill in missing values.

TABLE 2—GROWTH RATE EFFECTS OF DEMOGRAPHIC ADJUSTMENTS TO PRIVATE EMPLOYMENT, 1900–79

Period	Unadjusted Private Employment	Adjustment for Age	Additional Adjustment for Sex	Additional Adjustment for Immigration	Adjustment for Education	Quality Adjusted Private Employment
1900–79	1.40	0.06	−0.15	0.01	0.38	1.71
Major Periods:						
1900–29	1.77	0.13	−0.07	−0.03	0.09	1.88
1929–65	0.87	0.09	−0.19	0.06	0.64	1.47
1965–79	2.02	−0.18	−0.21	−0.02	0.34	1.95
Subperiods:						
1900–16	2.09	0.14	−0.06	−0.15	0.07	2.09
1916–29	1.38	0.11	−0.08	0.10	0.12	1.63
1929–48	0.88	0.15	−0.19	0.10	0.28	1.22
1948–65	0.86	0.03	−0.20	0.00	1.05	1.75
1965–73	1.84	−0.30	−0.17	−0.02	0.42	1.77
1973–79	2.27	−0.01	−0.28	−0.01	0.22	2.19

Note: Adjustments may not add due to rounding. Units: See Table 1.

TABLE 3—DIFFERENTIAL EFFECTS OF DEMOGRAPHIC ADJUSTMENTS TO PRIVATE EMPLOYMENT OVER MAJOR PERIODS, 1900–79

	Adjustment for				
Period	Age	Sex	Immigration	Education	Total
1900–29	0.07	0.08	−0.04	−0.29	−0.19
1929–65	0.03	−0.04	0.05	0.26	0.29
1965–79	−0.24	−0.06	−0.03	−0.04	−0.37

Note: Units: See Table 1.

Table 2 indicates the relative importance of the various demographic factors across the period by breaking the difference between unadjusted and quality-adjusted private-employment growth rates into age, sex, immigration, and education components.[14] Table 3 displays the differential effects in the three major periods by subtracting the average 1900–78 value of each adjustment from the value of the adjustment for the period. For example, failure to adjust for age in the productivity measures in Table 1 resulted in understating 1965–78 productivity growth by $-0.24-0.03=-0.27$ percent per annum relative to 1929–65. Overall, the productivity growth differential from 1929–65 was *overstated* by $0.29-(-0.19)=0.48$ percent per annum for 1900–29 and by $0.29-(-0.37)=0.66$ for 1965–78.

Table 4 illustrates that after private employment is adjusted for these demographic factors, both the hourly and employee productivity measures show *no* significant variation across the major periods. In particular, there is no indication of a secular productivity slowdown in 1965–78 vs. 1929–65 or indeed the entire twentieth century. Furthermore, as noted above, I have used the hours

[14] The growth-rate effect of education is separable from those of age, sex, and immigration, but the latter effects are not separable from each other. The age adjustment is most important for the recent periods so I compute first an age-adjusted private employment that makes no adjustment for sex or nativity. (Equations (2) and (3) are not used and $\alpha_1=\alpha_2=0.515$, $\alpha_3=1.0$.) Next an age-sex-adjusted private employment series is computed to find the marginal contribution of the sex adjustment. Finally, the marginal contribution of the immigration adjustment is calculated by comparing the *APE* derived using equations (1), (2), and (3) with this age-sex-adjusted private employment.

TABLE 4—AVERAGE ANNUAL GROWTH RATES OF PRODUCTIVITY MEASURES ADJUSTED FOR AGE, SEX, IMMIGRATION, AND EDUCATION, 1900–79

Period	Quality-Adjusted Private Employment (*QAPE*)	Average Hours Worked (*AHWP*)	Quality-Adjusted Private Hours Worked (*QATHWP*)	Gross Private Product (*GPP*)	Quality-Adjusted Hourly Productivity (*GPP/QATHWP*)	Quality-Adjusted Employee Productivity (*GPP/QAPE*)
1900–79	1.71	−0.23	1.48	3.23	1.75	1.52
Major Periods:						
1900–29	1.88	−0.22	1.66	3.42	1.76	1.54
1929–65	1.47	−0.27	1.20	2.98	1.78	1.51
1965–79	1.95	−0.12	1.83	3.48	1.65	1.53
Subperiods:						
1900–16	2.09	0.05	2.14	3.64	1.50	1.55
1916–29	1.63	−0.56	1.06	3.15	2.09	1.52
1929–48	1.22	−0.17	1.06	2.28	1.22	1.06
1948–65	1.75	−0.39	1.35	3.75	2.40	2.00
1965–73	1.77	−0.21	1.57	3.90	2.34	2.13
1973–79	2.19	0.00	2.19	2.91	0.72	0.72

measure most favorable to finding such a slowdown; if I had used alternative data on average private hours paid, average private nonsupervisory hours paid, or average civilian hours worked, the 1965–79 growth rates of output per quality-adjusted hour would have equalled or slightly exceeded the corresponding 1929–65 growth rate.[15]

Blau (1984), among others, has suggested that all or part of the sex differential in wages may reflect discrimination of the type in which women are paid less than their marginal product, presumably to compensate male employers or employees for associating with females. Fortunately that issue is not important for the question of whether there was a productivity slowdown: if women are weighted equally to their male counterparts of the same age ($\alpha_2 = 0.515$, $\alpha_3 = 1$), quality-adjusted private employment growth is 0.50 percent per annum higher in 1965–78 than in 1929–65 compared to a 0.48 percent per annum differential using our standard weights ($\alpha_2 = 0.42$, $\alpha_3 = 0.565$). Only for the 1900–29 period is the sex adjustment a substantial factor (0.12 percent per annum differential) relative to 1929–65. If we suppose that a half of the sex wage differentials reflect true marginal product differentials, the effect on the analysis *of productivity growth* is strictly *de minimis*.

In conclusion, changes in immigration laws and the entry of the baby boom divide the twentieth century into three major periods: before 1929, 1929–65, and after 1965. From the point of view of growth in employment and hours, each major period seems reasonably consistent, but the middle period is characterized by considerably lower growth than either of the exterior periods. Since the growth rate of gross private product declines by less than the decline in hours growth in the middle period, measured productivity growth rises. However, the demographic factors of age, sex, immigration, and education explain essentially all of the measured secular variation in hourly or per employee productivity growth. Thus, it appears that there is no substantial variation in trend private productivity growth over the twentieth century to be explained by variations in regulation growth, oil prices, the failure of American management, labor, or any of the other popular whipping boys. So far as broad

[15] The differences among the various measures of average hours may reflect differences in concepts—private vs. civilian, hours worked per employee vs. hours worked per employed person, hours worked vs. hours paid—or a yet-unidentified change in reporting procedures for the establishment data. Such a change occurred in 1934 with the introduction of the NIRA codes and minimum wages, but the hours data were adjusted for that as described in my 1984 study.

trends go, the U.S. productivity slowdown appears to be a case of statistical myopia.

Demographic adjustments also appear to explain observed variations in private employee productivity growth within the period 1900–29. The hourly productivity growth measure shows some residual variation (1.50 percent for 1900–16 vs. 2.09 percent for 1916–29), but this appears to be related to anomalous growth in our measure of average hours. It is left for economic historians to unravel whether the hourly productivity measure reflects a real phenomenon or simply measurement error.

Much more substantial variations in hourly and employee productivity growth are reported in Table 4 within the periods 1929–65 and 1965–78. For example, quality-adjusted hourly productivity growth is reported as 1.2, 2.4, 2.3, and 0.7 percent per annum for 1929–48, 1948–65, 1965–73, and 1973–78, respectively. It is the task of Section II to explain this residual variation.

II. Analysis of Intraperiod Variations

Section I argues that the quality-adjusted hourly productivity growth rate has had a constant secular value of 1.75 percent per annum throughout the twentieth century. Then how are we to explain the fact that this growth rate exceeded 2.3 percent per annum from 1948 to 1973 and was only 0.7 percent from 1973 to 1978? In this section it is first shown that the rapid 1948–65 growth is explained by the recovery of the capital-labor ratio from its abnormally low level at the end of World War II and the Great Depression. That is, the rapid (slow) growth in labor productivity during 1948–65 (1929–48) is due to abnormal movements in the capital stock relative to labor and output, and therefore is not reflected in *total* factor productivity growth or technical progress. Next it is demonstrated that the reported variations within the 1965–79 period appear to be the result of biases in measured output due to evasion of the 1974–74 price controls. Correction of these biases nearly eliminates any tendency for quality-adjusted hourly productivity growth to slow in 1973–79, or to be above 1.75 percent per annum in 1965–73.

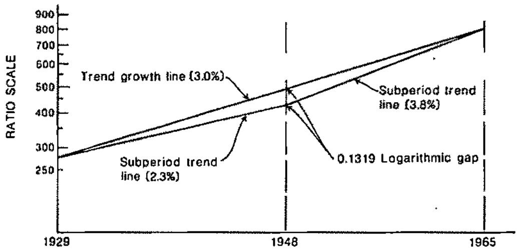


FIGURE 1. GROWTH TRENDS OF GROSS PRIVATE PRODUCT, 1929–65

A. *The 1929–65 Period*

The slow growth in 1929–48 and rapid catch-up growth in 1948–65 is attributed here to the very low ratio of investment to output during the Great Depression and World War II. The idea is that in 1948 we were quite poor in the sense of a low capital-labor ratio and it took until around 1965 to recover to the steady-state capital-labor and output-labor ratios as illustrated in Figure 1.

An implication of the approximately equal quality-adjusted hourly productivity growth rates for 1900–29, 1929–65, and 1965–79 is that the output-labor ratio is approximately the same in 1900, 1929, 1965, and 1978 after allowing for a constant rate of labor-augmenting technical progress.[16] Therefore, if capital growth explains the observed variation of output growth within the period, it follows that there was no significant intraperiod variation in technical progress (total factor productivity growth).

A simple and usually serviceable characterization of the aggregate production function is the Cobb-Douglas form

$$y = e^{\tau t} k^{\beta} l^{(1-\beta)}, \qquad (6)$$

where τ is the rate of total factor productivity growth and l is *measured* (quality-adjusted) labor input. This can equivalently

[16] As is well known, technical progress or total factor productivity growth is a euphemism for the increase in output which we cannot explain by the increase in *measured* inputs. Presumably, the constancy of its *average* growth rate over substantial periods reflects the law of large numbers and numerous independent contributing factors.

TABLE 5—IMPLICATIONS OF VARIATIONS IN CAPITAL AND QUALITY-ADJUSTED LABOR GROWTH FOR PRODUCTIVITY GROWTH, 1929–65

				Predicted $\Gamma(y/l)$			Actual
Period	γ	Γl	Γk	$\beta = 0.20$	$\beta = 0.25$	$\beta = 0.30$	$\Gamma(y/l)$
A: Hourly Productivity Concept							
1929–48	1.75	1.06	0.6	1.31	1.20	1.09	1.22
1948–65	1.75	1.35	4.0	1.93	1.98	2.02	2.40
Alternative Γk Estimates							
1948–65	1.75	1.35	5.0	2.13	2.23	2.32	2.40
1948–65	1.75	1.35	5.3	2.19	2.30	2.41	2.40
1948–65	1.75	1.35	5.6	2.25	2.38	2.50	2.40
B: Employee Productivity Concept							
1929–48	1.52	1.22	0.6	1.09	0.99	0.88	1.06
1948–65	1.52	1.75	4.0	1.67	1.70	1.74	2.00
Alternative Γk Estimates							
1948–65	1.52	1.75	5.0	1.87	1.95	2.04	2.00
1948–65	1.52	1.75	5.3	1.93	2.03	2.13	2.00
1948–65	1.52	1.75	5.6	1.99	2.10	2.22	2.00

Notes: Predicted $\Gamma(y/l)$ values are calculated using equation (8). Units: See Table 1.

be written in logarithmic form as

$$(7) \quad \log y = \beta \log k + (1-\beta)\log(le^{\gamma t}),$$

where $\gamma \equiv \tau/(1-\beta)$ is the constant rate of labor-augmenting technical progress. Subtracting $\log l$ from both sides of equation (7) and using Γ for the continuously compounded growth rate operator, we have

$$(8) \quad \Gamma(y/l) = \gamma + \beta[\Gamma k - (\Gamma(l+\gamma))].$$

That is, the growth rate of labor productivity equals the rate of technical progress plus the product of capital's share and the difference between the capital and the adjusted labor growth rates.

The capital growth rate has been estimated as about 0.6 and 4.0 percent per annum for 1929–48 and 1948–65, respectively.[17] Equation (8) is used to predict the observed quality-adjusted labor productivity growth rate for a capital share of 1/4 as well as alternative values of 0.2 and 0.3. Table 5 reports the results which indicate that the actual and predicted growth ratio of quality-adjusted hourly and employee productivity correspond quite closely for 1929–48. Thus the near cessation of investment during the Great Depression and World War II nicely explains the observed slowdown in productivity growth. Since the output-labor ratio has already been shown to return to its trend value by 1965, the solution seems to be complete.

Unfortunately, the second line of each part of Table 5 indicates that the predicted productivity growth falls short of actual growth by 0.3 or 0.4 percent per annum. It may be that this unexplained growth reflects a real temporary increase in technological progress that offsets an unusual fall in the capital-output ratio of some 24 percent over 1929–65,[18] but a simpler and economic explanation is also possible. Quite possibly the fault lies in the capital data themselves: the quantum leap in tax rates during World War II provides an incentive to write off as current

[17] Laurits Christensen and Dale Jorgenson (1978, pp. 35, 53) report series on corporate capital input and private domestic capital input with 1929–48 growth rates of 0.7 and 0.4 percent per annum, respectively. The NBER-Kendrick capital input series in U.S. Bureau of Economic Analysis (1973, pp. 192–93, Series A65) has an average growth rate of 0.6 percent per annum during this period. For 1948–65, Christensen and Jorgenson estimate that private domestic and corporate capital input grew at average rates of 4.0 and 3.8 percent per annum, respectively. The NBER-Kendrick data indicate only a 3.4 percent growth rate. The preference for the higher growth rate in the latter period is explained in the text below.

[18] The implied average growth rate of capital is 2.21 percent per annum while output grew at 2.98 percent; $\exp[(0.0221-0.0298)(36)] = e^{-0.2772} = 0.758$.

expense as much capital formation as possible; thus gross investment and capital growth could be systematically understated in the postwar period.[19] Suppose that a consistent data series would in fact show no decline in the capital-output ratio. This would imply that the true capital growth rate over 1948–65 is 5.6 percent per annum.[20] In order for this to be the case, firms would have had to alter their accounting practices so that reported net investment was reduced relative to earlier practices by almost 24 percent. Given the large incentives, this magnitude does not appear unreasonable, but further research is clearly indicated.[21] In any case, the lower halves of each part of Table 5 indicates that a 5.6 percent or even smaller capital growth rate would be sufficient to eliminate any apparent 1948–65 rise in total factor productivity growth.

In summary, measured capital growth variations can explain all of the 1929–48 slowdown in quality-adjusted labor productivity growth and a large part of the 1948–65 increase in that growth relative to trend. The remaining 0.3 to 0.4 percent excess growth in 1948–65 can be attributed either to an unexplained temporary increase in total factor productivity growth, or to changes in net investment reporting in response to increased income taxes.

B. *The 1965–79 Period*

It is not widely recognized that the main problem with productivity growth in 1973–79 is concentrated in the seven quarters 1973:II through 1974:IV:

> The productivity decline in 1973–74 was particularly striking. Labor productivity in the nonfarm business sector fell in every quarter from the second quarter of 1973 to the fourth quarter of 1974, dropping a total of 4.2 percent in a 7-quarter period. On the basis of the usual relationship between fluctuations in productivity and fluctuations in output, no more than 1 percentage point of that decline could be attributed to the sharp recession during the period. The additional drop of 3.2 percentage points accounts for much of the difference between the expected 2 percent annual growth rate between 1973 and 1977 and the 0.9 percent rate that actually occurred.[22]

This section will demonstrate not only that the progressive relaxation and ultimate removal of general price controls during 1973–74 can fully account for this anomalous excess productivity decline of 3.2 percent, but also that the imposition of these controls during 1971:II through 1973:I can account for the peculiarly rapid productivity growth observed during those quarters.[23] This rapid productivity growth permits us to reject the popular oil-price hypothesis in favor of the price control hypothesis.

Before examining the evidence, it is useful to sketch these two competing hypotheses. The oil-price hypothesis as developed by such authors as Robert Rasche and John Tatom (1977, 1981) asserts that higher oil prices will significantly lower the equilibrium level of output consistent with a given level of labor and capital and will further induce a fall over time in the level of capital. The price control hypothesis as developed in my 1976a,b studies asserts that measured real output was progressively overstated (and price understated) from the imposition of

[19] Overdeflation of gross investment due to undercorrection for quality changes would have a similar effect.

[20] That is, $[(19)(0.6\%)+(17)(5.6\%)]/36 = 2.96\% \approx \Gamma y = 2.98\%$.

[21] Obviously, I subscribe to the view that consistent data-collection procedures do not yield consistent data series when incentives or constraints change so as to alter the behavior of the optimizing agents who provide the data. This differs from the uncertainty principle in that economic analysis can be applied to estimate the nature of the changes.

[22] U.S. Council of Economic Advisers (1979, p. 70).

[23] See George Perry (1977, p. 37). In terms of my own short-run productivity growth function (equation (10) below), the residuals for these seven quarters are 0.0104, −0.0049, 0.0059, 0.0088, 0.0109, 0.0057, and 0.0107, respectively, for a sum of 0.0475. The residuals for the next seven quarters are −0.0074, −0.0071, −0.0040, −0.0087, −0.0121, −0.0075, and 0.0027, respectively, for a sum of −0.0441. The difference (0.003) is statistically insignificant and of the wrong sign for an oil-price effect on productivity.

price controls in 1971:III through 1973:I and that this overstatement was progressively eliminated under Phase III and decontrol (1973:II–1974:III). Sung Hee Jwa (1983) has extended my basic model by a formal analysis of firm and industry equilibrium. My 1982 article examines the oil-price and price control hypotheses in detail using both U.S. and international data and finds that the preponderance of evidence supports the price control hypothesis. This evidence will be supplemented below by directly estimating a productivity growth equation and by other empirical evidence. The second oil shock can be included and quarterly data used only at the cost of not making the demographic adjustments of Section I. Instead it will be assumed that from 1966 onwards demographic factors reduce conventionally measured labor productivity growth by a constant amount.

First we wish to test whether oil prices, price controls, or both had a significant influence on productivity growth other than via any temporary effects causing unemployment and employment to differ from their steady-state values. Standard productivity equations have deflated values on both sides inducing spurious correlation if the price control hypothesis is true. Fortunately, a simple dynamic Okun's Law extended by other current and leading labor market indicators provides very respectable explanatory power without potential spurious correlation. The basic equation used corresponds (with one minor exception) in right-hand variables to equation (5) in my 1982 article:[24]

$$(9)\quad \Delta\log(y/l)_t = a_1 + a_2TS_t + a_3\Delta u_t + a_4\Delta\log E_t + a_5\Delta\log E_{t+1} + \varepsilon_t,$$

where $(y/l)_t$ is the private-hours-paid definition of labor productivity,[25] TS_t is a time shift dummy equal to 0 before 1966 and 1 otherwise,[26] u_t is the total unemployment rate, and E_t is employment in manufacturing, mining, and construction. Note that the cyclical indicators used in this equation are all based on counts of individuals and so not subject to possible reporting biases (as are deflated series) under price controls. The estimated equation for 1950:II–1983:I is[27]

$$(10)\quad \Delta\log(y/l)_t = \underset{(7.65)}{0.0069} \underset{(-2.21)}{-0.0027TS_t} \underset{(-3.47)}{-0.010\,\Delta u_t} \underset{(0.44)}{+0.039\Delta\log E_t} \underset{-(5.62)}{-0.293\Delta\log E_{t-1}}$$

$$S.E.E. = 0.00690,\ \bar{R}^2 = 0.28,\ D\text{-}W = 1.95.$$

This equation does reasonably well at explaining quarterly fluctuations in productivity growth, although only the current change in unemployment and the lagged growth rate of employment are significant among the cyclical indicators.

To test the price control hypothesis, a simple quantitative variable was formed: CD_t grows linearly from 0 in 1971:II to 1 in 1973:I and then falls linearly back to 0 in 1974:IV. The deflated dollar price of a barrel of Venezuelan oil was used for the oil price P_t. The following general equation was estimated with up to a one-year adjustment lag permitted for oil prices to take effect:

$$(11)\quad \Delta\log(y/l)_t = a_1 + a_2TS_t + a_3\Delta u_t + a_4\Delta\log E_t + a_5\Delta\log E_{t-1} + a_6\Delta CD_t + \sum_{i=0}^{3} a_{7+i}\Delta\log P_{t-i} + \varepsilon_t.$$

[24] The minor exception is that the layoff rate was dropped because the Bureau of Labor Statistics stopped collecting the data in 1981. The layoff rate was marginally significant in my 1982 real-income equation but had a t-statistic of only -0.4 over the available observations when added to the present equation (9). Dropping the layoff rate did not cause any significant change in the remaining coefficients or their t-statistics.

[25] This was the most convenient measure of private productivity available quarterly. All data for this quarterly analysis were taken from the Citibase data bank.

[26] This variable is supposed to capture the differential effects of the demographic adjustments summarized in Tables 2 and 3 above.

[27] The t-statistics are shown in parentheses below the estimated coefficients. The estimation is over the entire period for which data were available on Citibase at the time the final draft of this paper was prepared.

TABLE 6—TEST STATISTICS FOR ALTERNATIVE VERSIONS OF EQUATION (11)

Line Number	Restrictions	*S.E.E.*	Value and *t*-Statistic for		*F*-Statistic for	
			a_6	a_7	$a_7=a_8=a_9=a_{10}=0$	$a_8=a_9=a_{10}=0$
1	none	0.00663	0.04104 (2.69)[c]	−0.00413 (−0.62)	$F(4,122)=0.42$	$F(3,122)=0.49$
2	$a_8=a_9=a_{10}=0$	0.00659	0.04515 (3.46)[c]	−0.00301 (−0.48)	–	–
3	$a_7=a_8=a_9=a_{10}=0$	0.00657	0.04691 (3.77)[c]	–	–	–
4	$a_6=0$	0.00679	–	−0.00868 (−1.31)[a]	$F(4,123)=2.00$	$F(3,123)=1.94$
5	$a_6=a_8=a_9=a_{10}=0$	0.00687	–	−0.00928 (−1.47)[a]		
6	$a_6=a_7=a_8=a_9=a_{10}=0$	0.00690	–	–	–	–

[a] Significant at 10 percent level or better.
[b] Significant at 5 percent level or better.
[c] Significant at 1 percent level or better.

Table 6 reports the results of various alternative hypothesis tests that might be conducted. Line 1 pertains to equation (11) as stated while all the other lines involve various zero constraints on $a_6,\dots,a_{10}$.[28] We see that whenever the price control variable is included it is significant at the 1 percent level or better. The oil variables, in contrast, are never significant except for lines 5 and 6 in which, with the price control variable forced out, current oil-price growth is significant at the 10 percent level on a one-tailed test. I conclude that oil-price changes had no significant effect on productivity growth. Note particularly that a major increase in real oil prices occurred between 1979:I and 1980:I, but no direct effect was detectable.

The final form of the regression is

$$\begin{aligned}(12)\qquad \Delta\log(y/l)_t &= \underset{(8.06)}{0.0070} \;-\underset{(-2.39)}{0.0027TS_t}\\ &\quad -\underset{(-3.42)}{0.010\Delta u_t}+\underset{(0.31)}{0.027\Delta\log E_t}\\ &\quad -\underset{(-5.61)}{0.280\Delta\log E_{t-1}}+\underset{(3.77)}{0.04691\Delta CD_t}\end{aligned}$$

$$S.E.E.=0.00657,\ \bar{R}^2=0.35,\ D\text{-}W=2.08.$$

Consider the implications of this equation for the level of productivity in the year 1973. The average value of CD_t in 1973 is 0.7857 which, when multiplied by 0.04691, implies that the logarithm of labor productivity in 1973 was overstated by 0.0369.[29] This means that the 1965–73 growth rates of private labor productivity are *overstated* by 3.69/8 = 0.46 percent per annum and correspondingly that the 1973–79 growth rates are understated by 3.69/6 = 0.61 percent per annum. Table 7 shows that applying this correction to the quality-adjusted productivity growth rates of Section I eliminates any evidence of a major 1973–79 productivity slowdown. Instead the picture is one of remarkably stable productivity growth over the period 1965–79 after accounting for the 1973 measurement biases. The growth rate of output per quality-adjusted employee remains within 0.20 percent per annum of the century average rate of 1.52 percent per annum. The same nearly holds for output per quality-adjusted hour worked, and would hold on any of the alternative measures of hours.[30]

[28] Thus line 6 refers to equation (10) as reported above.

[29] This estimate is an estimate of the overstatement in *deflated* private output and thus applies equally well to the annual quality-adjusted private productivity measures reported in Table 4 above.

[30] Recall that the hours series used showed the fastest relative growth in hours in 1973–79 among the alternatives. The century average growth in *GPP/QATHWP* was 1.75 percent per annum.

TABLE 7—CALCULATION OF QUALITY-ADJUSTED LABOR PRODUCTIVITY GROWTH RATES ADJUSTED FOR PRICE CONTROL REPORTING BIASES, 1965–78

	Productivity Measures	
	Hourly Productivity (*GPP/QATHWP*)	Employee Productivity (*GPP/QAPE*)
Reported Growth, 1965–73	2.34	2.13
Less (0.0369/8)×100	−0.46	−0.46
Adjusted Growth, 1965–73	1.88	1.67
Reported Growth, 1973–79	0.72	0.72
Plus (0.0369/6)×100	+0.61	+0.61
Adjusted Growth, 1973–79	1.33	1.33

Note: Units: See Table 1.

The remaining small fluctuations of the productivity growth rates around secular trends can be reasonably attributed to small sample fluctuations and the effect of small differences in unemployment rates for 1965, 1973, and 1979.

It is of course true that price controls could have had real effects, but these effects should have operated by changing unemployment and employment. The estimated coefficient of ΔCD captures some additional impact which must either measure output overstatement or some shift in the relationship of output to labor inputs for a reason yet to be proposed in the literature.

C. *Additional Discussion and Evidence on the Price Control Hypothesis*

There are three popular models of the effects of President Nixon's Economic Stabilization Program (ESP). The first, used as the basic economic support for the program, argued that sticky expectations and nominal contracts would delay adjustment to a new lower, noninflationary equilibrium. The ESP, the argument goes, would accelerate the adjustment process and minimize the transitional increase in unemployment. The second view, associated with Robert Barro and Herschel Grossman (1974) and Paul Evans (1982), argues that general price controls reduce real output and inflation by inducing increased consumption of leisure. The third view, which I have proposed (1976a, b; 1982) argues that the ESP was largely window dressing and was easily evaded by minor covert quality depreciation both in physical products and services and in the terms on which they were sold.

Needless to say, these views are not mutually exclusive. For example, the ESP most probably reduced the unemployment associated with the existing macroeconomic conditions so that true output increased while reported output increased even more due to covert quality depreciation not captured in the official price indices. For the analysis of productivity growth, we are interested in the reporting effects and not the real effects.[31]

It may be useful to look more closely at how these reporting effects could occur. Recall that the ESP established price controls relative to the base-period price of each product produced by each firm. New,

[31] Thus equation (12) in the text above is a way of estimating the spurious increase in reported output conditional upon whatever *real* effect on unemployment and employment may have occurred. The reporting hypothesis implies that the official division of nominal amounts into quantities and prices is generally suspect during 1971–74. Deriving implications from (possibly) incorrectly deflated data is a task not unlike that of George Smiley in LeCarre's *Tinker, Tailor, Soldier, Spy*. Frequently the data seem to tell one consistent story if they are taken at face value and another consistent story if they are assumed biased by the price control program. The challenge is to find cases in which only one of the hypotheses fit.

higher-quality products could be introduced at higher prices reflecting their higher costs. During Phases 1 and 2 (August 1971–January 1973) the controlled prices generally fell relative to the prices which otherwise would have prevailed. This provided an increasing incentive to make covert quality depreciations in existing goods and to claim spurious quality appreciations in new goods. Or to say the same thing, there was an increasing incentive to publicize every quality improvement and to shade the quality of existing products. If firms reacted to these incentives as we normally suppose, then those collecting data for computing price indices would likely miss more quality depreciation and record more quality appreciation than normal. Controls become progressively less binding under the subsequent Phases III and IV ending *de jure* on April 30, 1974, with the expiration of legislative authority and *de facto* in the third quarter with the expiration of certain pricing agreements negotiated in exchange for early decontrol.[32] So during this period firms had an incentive to progressively restore their products to their nominal quality. To the extent price data collectors missed the shading of quality during Phases I and II, they should equally have missed its restoration during the relaxation and removal of controls.

Before going any further, we must consider whether this story is empirically plausible. Some economists, especially those responsible for collecting the price data, have doubted that any significant quality shading could have been missed. However, the price control hypothesis does not require any huge errors. The estimate in equation (12) of the missed quality decrease—or better, of the decrease in the quality improvement which was missed—only amounts to about 0.2 percent per month (2.7 percent per annum). This magnitude is very small not only in absolute terms, but also relative to the supposed margin of error in quality adjustments.[33] Missed quality change always imparts some bias to measured real *GNP* growth, but price controls impart incentives which change the bias in predictable ways.

Another possible objection is that firms shading quality would be caught by the IRS's monitoring of the profit margin ceiling. But this is not the case in a balanced inflation in which prices, costs, sales, and profits all rise in proportion. Everyone can accurately report the dollar amounts of revenues, costs, and profits since the profit margin ceiling was purely window dressing absent any effective controls on (quality-adjusted) costs. Thus nominal value-added and nominal *GNP* will be correctly computed; only its division between real *GNP* and the deflator will be biased.

Above (and in my 1982 article) I have already shown that misreporting under price controls can explain the anomalous behavior of Okun's Law during the price control period.[34] Okun's Law should underpredict

[32] Charles Cox (1980) uses October 1973 (i.e., 1973 : IV instead of 1973 : II) as the beginning of the decontrol period since that was the beginning of sector-by-sector decontrol. I believe that the Phase III removal of requirements for prior approval of price increases was a major relaxation since, as explained below, the remaining profit margin ceiling was consistent with any rate of inflation. Some macroeconomic evidence supporting 1973 : II instead of 1973 : IV as the start of decontrol is offered below, but the issue is not crucial. The same evidence is consistent with Cox's view that the controls had no effect on price levels past 1974 : II (as compared to my 1974 : III) nor on growth rates past 1974 : III.

[33] Economists have traditionally argued that missed quality improvements might imply a 2 percent measured inflation even if the "true" price level was constant. See, for example, Gardner Ackley (1961, p. 87), Price Statistics Review Committee (1961, pp. 35–39), and Zvi Griliches (1961).

[34] In a simple dynamic Okun's Law regression, one obtains

$$\underset{(14.58)}{\Delta \log y = 0.0086} \; \underset{(-13.49)}{-0.0185\Delta u} \underset{(2.94)}{+0.0377\Delta CD},$$

$$S.E.E. = 0.00678,\ \bar{R}^2 = 0.61,\ D\text{-}W = 2.02,$$

$$\text{Period} = 1950:\text{II}\text{–}1983:\text{I}.$$

The coefficient on ΔCD is within one standard error of the estimate of 0.04691 obtained in the productivity-growth equation (12) in the text. Note, that although Okun's Law is sometimes reversed to explain Δu given $\Delta \log y$, that form is not appropriate to the current case in which relatively large measurement error is hypothesized for $\Delta \log y$ as compared to Δu.

output growth from 1971 : III through 1973 : I when the growth is overreported and correspondingly overpredict output growth during the decontrol period. For 1971 : II to 1974 : IV as a whole, Okun's Law predicts total growth in real *GNP* rather well.[35]

Let us see what other evidence can be offered in support of the reporting hypothesis. A simple check on the hypothesis that the nominal *GNP* data will be unaffected involves running a simple reduced-form regression explaining nominal *GNP* (Y) growth by a distributed lag on nominal money ($M1$) growth and ΔCD:

$$(13)\quad \Delta \log Y = h_0 + \sum_{i=0}^{7} h_{1+i} \Delta \log M1B_{t-i} + h_9 \Delta CD.$$

The estimated regression can be summarized as

$$(14)\quad \Delta \log Y = \underset{(3.33)}{0.0100} + \sum_{i=0}^{7} h_{1+i} \Delta \log M1B_{t-i} + \underset{(0.07)}{0.0012} \Delta CD$$

$$\sum_{i=0}^{7} h_{1+i} = 0.7650$$

$$S.E.E. = 0.00897,\ \bar{R}^2 = 0.19,\ D\text{-}W = 1.80,$$

$$\text{Period} = 1961 : \text{I–}1983 : \text{I}.$$

Thus controls do not appear to have any significant impact on nominal *GNP*.[36] An analogous regression confirms the hypothesized negative impact of controls on the inflation rate as measured by the *GNP* deflator *PD*:[37]

$$(15)\quad \Delta \log PD = \underset{(-0.11)}{-0.0003} + \sum_{i=0}^{7} k_{1+i} \Delta \log M1B_{t-i} - \underset{(-2.69)}{0.0300} \Delta CD$$

$$\sum_{i=0}^{7} k_{1+i} = 0.9892,$$

$$S.E.E. = 0.00389,\ \bar{R}^2 = 0.29,\ D\text{-}W = 2.12,$$

$$\text{Period} = 1961 : \text{I–}1983 : \text{I}.$$

Thus regression analysis of U.S. data on real *GNP*, nominal *GNP*, and the *GNP* deflator indicates that price controls had no effect on reported total nominal spending, but only upon its division into prices and output. The fact that real output and productivity growth appear to rise and fall relative to that predicted by labor market conditions strongly supports the reporting hypothesis.

Separate evidence in support of the reporting hypothesis is to be found in comparisons of reported deflated *GNP* not with inputs but with alternative measures of output. George Terborgh (1979) has noted the anomalous behavior of reported real *GNP* relative to the Federal Reserve index of manufacturing production in the period 1971–74. As noted by Terborgh, the FRB index is based primarily on counts of physical units. Terborgh shows that although normally the FRB index grows faster than real *GNP*, this is not true in 1971 and 1972. Furthermore, measured real *GNP* falls sharply relative to the FRB index in 1973, 1974, and 1975.[38] A formal check on whether or not price controls move measured real *GNP*

[35] See my 1976a article.

[36] These results are inconsistent with the Barro-Grossman-Evans view discussed above. The same qualitative results for price controls (i.e., no effect) are obtained if the estimation period is 1961 : I–1980 : IV, but the $\bar{R}^2$ is considerably higher, *S.E.E.* lower, *D-W* closer to 2, and Σh_i closer to 1. This may be suggestive evidence that shifts in money demand associated with the 1981 introduction of nationwide NOW accounts and other recent reforms have disturbed the nominal-income equation, but that debate is beyond the scope of this paper.

[37] This equation was estimated with a first-order correction for autocorrelation ($\hat{\rho} = 0.5391$). Without this correction the coefficient of ΔCD was estimated as -0.0373 (standard error 0.0089, *t*-statistic -4.19) and the *S.E.E.* was 0.0045 with a *D-W* of 1.01.

[38] Terborgh's use of annual data spreads the adjustment period into 1975 since the *average* 1974 real *GNP* data will be overstated on the price control hypothesis.

compared to what would be expected from the Index of Manufacturing Production (*IMP*) involves running the regression:

$$(16)\quad \Delta \log y = \underset{(9.55)}{0.0053} + \underset{(16.64)}{0.3166}\Delta \log IMP + \underset{(2.62)}{0.0312}\Delta CD,$$

$$S.E.E. = 0.00630,\ \bar{R}^2 = 0.69,\ D\text{-}W = 2.08,\quad \text{Period} = 1948\!:\!\text{II}–1983\!:\!\text{I}.$$

Note that the coefficient on ΔCD is some 0.016 smaller than that estimated in above for the productivity growth equation (12). Although the difference is not statistically significant, it is to be expected since *IMP* includes some deflated as well as physical unit series.[39] It is proposed in future research to follow up these very promising results by using the underlying individual data series on physical units of homogeneous commodities to construct an independent estimate of real *GNP* for analysis of recent productivity growth.

In summary, there is a considerable body of evidence that the uneven productivity growth reported in 1965–79 can be explained by reporting biases in 1971–74 and normal cyclical factors. My 1982 article showed that similar adjustments may be required in those countries which adopted programs modeled on the ESP during 1971–74. These results support the basic conclusion of this paper: that there have been no substantial variations in secular U.S. labor productivity growth after adjustment for changing demographic trends.

III. Conclusions and Areas for Future Research

The results of this study can be clearly summarized by the use of two figures. Figure 2 illustrates the logarithm of hourly productivity measured in the standard way by *GPP/THWP*. It is difficult if not impossible to discern any overall trend although 1900–29 and 1929–48 might be identified as periods of slow growth followed by rapid growth during 1948–65 and then slowing growth over 1965–73 and 1973–79. The logarithm of quality-adjusted hourly productivity (*GPP/QATHWP*) is plotted in Figure 3. Here a constant trend line dominates the data except during the Great Depression and Korean War eras of slow investment and subsequent rapid recovery. With demographic factors accounted for, the anomalous productivity gains in 1972 and 1973 (which I attribute to measurement biases) stick out like the proverbial sore thumb.

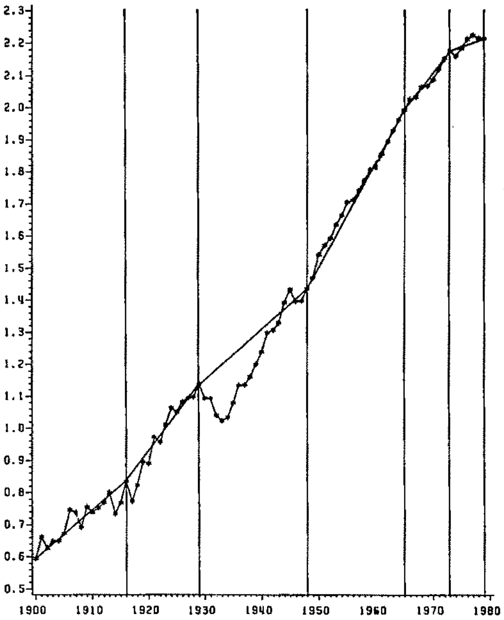


FIGURE 2. LOGARITHM OF HOURLY PRODUCTIVITY, log(*GPP/THWP*)

The major conclusion to be drawn is that there have been no substantial variations in trend growth rates of private labor productivity since 1900 if reasonable adjustments are made for the effects of demographic trends on the average quality of labor. Even if one were to ignore the effects of demo-

[39] A formal *F*-test was conducted for both this regression and the one reported in fn. 34 above to test the implicit hypothesis that the coefficient was the same during the decontrol and control periods. The hypothesis was not rejected.

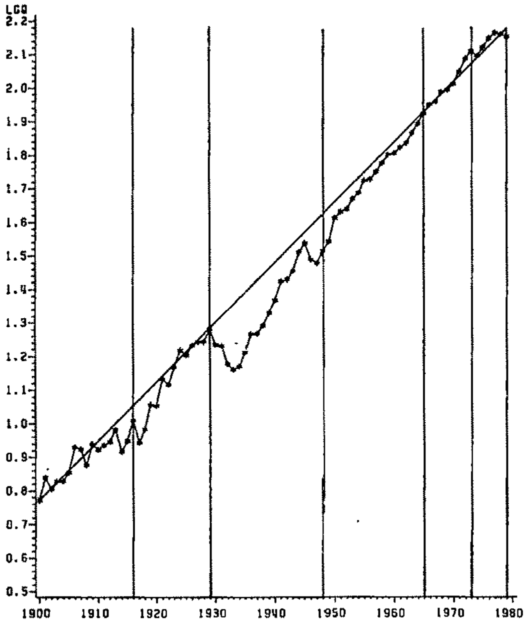


FIGURE 3. LOGARITHM OF QUALITY-ADJUSTED HOURLY PRODUCTIVITY, 1900–79, log(*GPP*/*QATHWP*)

graphic shifts, the measured growth rates of productivity, total private hours, and private employment have essentially the same values in 1900–29 as in 1965–79 so that panic may be premature.

The slow labor productivity growth in 1929–48 can be explained by the near cessation of capital formation, but measured increases in capital growth in 1948–65 are too small to fully account for the catch-up of labor productivity. Further research is required to determine whether this is due to problems in the measurement of capital or to other yet undiscovered factors. The slowdown in productivity growth within the period 1965–79 can be explained by measurement biases induced by evasion of price controls. Increased oil prices do not play a significant role.

Taken as a whole, the evidence does not support the view that there has been a substantial, inexplicable decline in total factor productivity growth since 1965 and especially since 1973. Instead the evidence presented here indicates that there has been a surprisingly stable growth rate of total factor productivity throughout the twentieth century. Only in 1948–65 is there any evidence of a substantial (0.2 to 0.4 percent per annum) temporary increase in total factor productivity growth and there are good economic reasons to suspect that this may be an artifact of tax-induced changes in accounting procedures.

One may ask why this study has succeeded in finding a constant underlying trend in appropriately measured labor (or total factor) productivity where others have been unable to do so. That may be like asking why Sally solved Rubik's Cube fastest, but it is a question to which I must respond. In part, the analysis in this article starts from a firmly macroeconomic viewpoint: this excludes worrying about a lot of trees which may obscure the forests. For example, Denison (1974) attempts to measure an effect on productivity from shifting a *given* worker from one industry to another; here the assumption is that the allocation of resources will assure a normally efficient use of the pool of human and produced capital. Further, as a macroeconomist I was already aware of the price control biases during 1971–74 and did not try to fashion an explanation to fit distorted data. In my 1976c article, I had shown that the change in reporting procedures for hours the NRA codes and hours laws became effective significantly distorted hourly wage data; so correcting the discontinuity in hours data at 1933 was an obvious improvement over previous work.

Thus, the major difference between this article and the previous literature seems to be in the conceptual approach. Here I start with an aggregate production function and try to measure aggregate labor very carefully since capital will conform to labor (absent a Great Depression and World War!). The major quantifiable factors (besides numbers of workers) which seem important are age, sex, education, and immigration as determinants of quality of the labor force and the average number of hours worked. The fact is that this approach seems to work where previous attempts based on growth accounting have not.

A considerable program for future research which has been noted in previous sections can be briefly summarized. 1) An interesting issue for economic historians is whether the intraperiod inconsistency between hourly and employee productivity growth for 1900–29 reflects a real phenomenon or indicates a measurement problem in the data on average hours worked. 2) Certainly improvements can be made to the demographic adjustments reported here. Doubtless others will test these results by doing so. 3) The failure of the closed-economy neoclassical growth model suggests an (industrialized) world linked by capital flows which are quite responsive in the long run. Similar analyses of other economies could similarly explain their postwar labor productivity recoveries and slowdowns. It further suggests investment incentives may be more effective than saving incentives as means of increasing domestic capital stock. 4) In light of the discrepancy between full recovery in the output-labor ratio and incomplete recovery in the capital-labor ratio, a reexamination of the consistency of existing estimates of net investment and the capital stock is in order. 5) The potential importance of price control induced biases in deflated (and deflator) data during the 1971–74 period is once again demonstrated. Further evidence is called for on this issue, but the time has come to reexamine various claims for effects of oil-price and other variables which may serve as a proxy for these reporting biases.

A warning note is in order. The fact that factors such as regulation, governmental size, oil prices, management practices, educational quality, moral fiber, and the like have not been required to fully explain twentieth-century variations in labor productivity does *not* imply that they have been unimportant. Any or all of them may have been quite important in determining the trend value of total factor productivity growth. Nonetheless in the aggregate their impact has caused quality-adjusted total factor productivity growth to evolve as if following a random walk with constant drift and small variance. For this type of process the average growth rate of total factor productivity growth converges over considerable periods to the constant drift.

Fortunately, the analysis leaves considerable room for optimism. The major factor reducing productivity growth over 1965–79 was the increasingly youthful labor force. Now as the smaller post-pill generation enters the labor force and the baby-boom generation ages, this effect will be operating in the opposite direction to increase output per (unadjusted) hour. The acculturation of recent immigrants will be another positive, albeit much smaller, factor. The sex factor may cease to slow productivity growth for either of two reasons: 1) female participation rates are approaching those of men for younger women, so the disproportionate growth of women workers may soon slow,[40] 2) as lifetime market work becomes the norm for women, their investment in human capital should rise toward that of men. The one factor which public policy can directly influence is education. Whether the last two decades of the twentieth century will witness the equivalent of another GI Bill and whether the same marginal effects would continue to accrue is an open question, but surely three positive factors are enough for the dismal science.

Data Appendix

The primary data base for this paper is reported in my 1984 study. Table A1 extracts the data on private employment (PE), total hours worked in the private sector ($THWP$), the implicit average hours worked in the private sector ($AHWP$), and gross private product (GPP) from that source. Table A2 reports the data for the five main calculated series: quality-adjusted private employment ($QAPE$), quality-adjusted total private hours worked ($QATHWP$), the average years since migration of foreign-born workers by sex (Z_M and Z_F), and the median years of education E.

[40] See my 1984 study, ch. 2.

TABLE A1—PRIVATE EMPLOYMENT, HOURS, AND OUTPUT 1900–79

Year	*PE*	*THWP*	*AHWP*	*GPP*	Year	*PE*	*THWP*	*AHWP*	*GPP*
1900	26692.1	57.242	118.371	103.680	1940	43318.0	87.300	111.240	301.240
1901	27679.7	59.715	119.079	115.835	1941	45690.0	94.700	114.405	346.660
1902	28503.1	62.371	120.783	116.860	1942	48267.0	102.100	116.759	376.989
1903	29173.3	64.111	121.301	122.717	1943	48390.0	105.800	120.683	399.700
1904	29389.2	63.195	118.689	120.960	1944	47917.0	104.400	120.261	420.358
1905	30535.2	66.309	119.864	130.039	1945	46876.0	98.700	116.220	413.912
1906	32257.1	68.965	118.010	145.416	1946	49655.0	99.700	110.828	403.069
1907	32813.6	70.522	118.628	147.612	1947	51564.0	101.800	108.972	412.300
1908	31618.7	67.408	117.675	134.725	1948	52693.0	102.600	107.476	431.500
1909	33376.9	71.163	117.686	151.566	1949	51795.0	98.800	105.289	429.825
1910	33993.3	72.995	118.527	153.031	1950	52892.0	100.500	104.880	470.050
1911	34364.6	74.369	119.452	158.010	1951	53572.0	104.000	107.155	500.375
1912	35570.0	76.659	118.958	165.478	1952	53641.0	104.700	107.737	515.275
1913	36386.6	77.300	117.260	172.068	1953	54534.0	104.900	106.175	538.550
1914	35589.3	75.834	117.615	158.010	1954	53358.0	100.600	104.067	531.825
1915	35477.5	75.193	116.988	162.256	1955	55256.0	104.100	103.989	572.925
1916	37268.7	80.597	119.368	185.687	1956	56521.0	105.600	103.126	584.85
1917	37350.6	82.337	121.678	178.365	1957	56455.0	104.200	101.878	594.70
1918	37265.9	81.604	120.870	185.833	1958	55197.0	100.000	100.000	590.60
1919	37906.7	79.040	115.092	193.448	1959	56547.0	103.300	100.834	629.52
1920	38044.5	80.047	116.137	195.205	1960	57425.0	104.500	100.446	641.95
1921	35805.3	72.079	111.116	190.666	1961	57152.0	102.800	99.284	657.75
1922	38402.7	77.483	111.368	201.942	1962	57812.0	104.800	100.060	697.77
1923	41176.6	83.619	112.091	229.765	1963	58537.0	105.600	99.575	727.35
1924	40703.8	81.696	110.785	236.648	1964	59709.0	107.700	99.561	767.55
1925	42297.3	84.627	110.436	242.066	1965	61014.0	111.100	100.508	816.57
1926	43355.5	87.557	111.472	258.321	1966	62111.0	113.906	101.226	863.97
1927	43292.4	87.099	111.050	260.518	1967	62981.0	115.704	101.404	883.70
1928	43419.6	87.832	111.656	263.447	1968	64081.0	117.416	101.137	925.65
1929	44565.0	89.572	110.942	279.702	1969	65707.0	120.250	101.015	952.00
1930	42332.0	83.436	108.793	249.226	1970	66073.0	117.680	98.309	949.50
1931	39136.0	76.475	107.860	228.274	1971	66239.0	118.227	98.519	985.67
1932	35715.0	67.775	104.745	191.938	1972	68368.0	121.801	98.336	1048.10
1933	35594.0	67.042	103.964	186.810	1973	70677.0	126.573	98.850	1115.90
1934	37591.0	72.000	105.722	202.634	1974	71765.0	127.448	98.025	1105.65
1935	38779.0	75.900	108.034	223.293	1975	70097.0	122.619	96.555	1088.97
1936	40742.0	81.600	110.551	253.622	1976	72614.0	126.236	95.957	1154.12
1937	42544.0	86.700	112.485	270.032	1977	75419.0	132.075	96.662	1223.27
1938	40337.0	79.200	108.377	252.889	1978	78701.0	140.018	98.202	1285.05
1939	41755.0	83.300	110.116	276.478	1979	80998.0	145.081	98.867	1329.15

Note: Private Employement (*PE*); Total Hours (*THWP*); Average Hours (*AHWP*); Gross Product (*GPP*).
Sources: PE, THWP, and *GPP* are from my 1984 study, Tables A7, A19, and A20, respectively; $AHWP = (THWP/PE) \times (PE_{1958})$ so that the base year value (1958 = 100) of the *THWP* index is preserved.

TABLE A2—QUALITY-ADJUSTED PRIVATE EMPLOYMENT AND HOURS, YEARS SINCE MIGRATION, AND MEDIAN EDUCATION, 1900–79

	Quality-Adjusted		Years since Migration		
Year	Private Employment (*QAPE*)	Private Hours (*QATHWP*)	Males (Z_M)	Females (Z_F)	Median Education (*E*)
1900	36435.7	47.998	22.6674	7.3866	8.0000
1901	37794.0	50.085	21.9248	7.7412	8.0099
1902	38887.9	52.272	20.7291	8.0444	8.0199
1903	39731.8	53.635	19.1892	8.2398	8.0299
1904	39996.8	52.831	18.0640	8.4023	8.0399

TABLE A2—(*Continued*)

Year	Quality-Adjusted Private Employment (*QAPE*)	Quality-Adjusted Private Hours (*QATHWP*)	Years since Migration Males (Z_M)	Years since Migration Females (Z_F)	Median Education (*E*)
1905	41479.7	55.332	16.6296	8.5072	8.0498
1906	43745.6	57.452	15.3560	8.5674	8.0599
1907	44390.2	58.603	13.9928	8.6135	8.0699
1908	42834.6	56.096	13.6240	8.7904	8.0799
1909	45280.6	59.304	13.2629	9.0324	8.0899
1910	46115.5	60.829	12.5829	9.1581	8.1000
1911	46636.8	61.997	12.2705	9.2848	8.1099
1912	48572.3	64.303	12.0623	9.4171	8.1999
1913	49374.1	64.432	11.5134	9.4377	8.1299
1914	48265.0	63.175	11.0806	9.4358	8.1399
1915	48288.5	62.868	11.4446	9.8364	8.1498
1916	50903.2	67.621	11.8014	10.2617	8.1599
1917	51275.4	69.434	12.1555	10.6709	8.1699
1918	51690.0	69.530	12.6642	11.1911	8.1799
1919	52552.2	67.311	13.1303	11.6892	8.1899
1920	52686.1	68.095	13.3064	11.9479	8.2000
1921	49628.9	61.371	13.1465	11.8868	8.2198
1922	53351.6	66.123	13.4958	12.1935	8.2396
1923	57304.8	71.484	13.5692	12.3911	8.2595
1924	56717.5	69.927	13.4568	12.4623	8.2794
1925	59097.5	72.632	13.7836	12.8174	8.2994
1926	60734.1	75.343	14.0887	13.1591	8.3194
1927	60792.2	75.131	14.3434	13.4793	8.3395
1928	61131.2	75.962	14.6407	13.7915	8.3596
1929	62905.7	77.667	14.9736	14.1043	8.3798
1930	59921.4	72.549	15.3459	14.4371	8.4000
1931	55467.4	66.580	15.8651	14.9130	8.4198
1932	50688.4	59.087	16.4378	15.4647	8.4396
1933	50579.2	58.520	17.0194	16.0329	8.4595
1934	53479.3	62.921	17.5927	16.5898	8.4794
1935	55228.8	66.401	18.1593	17.1342	8.4994
1936	58084.5	71.462	18.7217	17.6737	8.5194
1937	60710.0	75.999	19.2620	18.1882	8.5395
1938	57610.1	69.484	19.7739	18.6684	8.5596
1939	59679.1	73.135	20.2515	19.1243	8.5798
1940	61977.7	76.727	20.7432	19.5963	8.6000
1941	65567.7	83.480	21.2646	20.0962	8.6676
1942	69423.1	90.208	21.8240	20.6353	8.7357
1943	69413.7	93.226	22.3895	21.1831	8.8043
1944	69286.4	92.731	22.9470	21.7151	8.8734
1945	68669.9	88.817	23.4941	22.2115	8.9432
1946	73541.0	90.704	23.9784	22.4542	9.0134
1947	77447.4	93.923	24.3396	22.6387	9.0842
1948	79364.1	94.926	24.6321	22.7764	9.1556
1949	78480.9	91.960	24.8572	22.8908	9.2275
1950	80406.3	93.849	24.8947	22.9062	9.3000
1951	82230.8	98.061	25.0187	23.0240	9.4225
1952	83413.6	100.012	25.0171	22.9757	9.5466
1953	85932.7	101.539	25.2558	23.1282	9.6723
1954	84990.3	98.431	25.3708	23.2031	9.8000
1955	88311.3	102.200	25.3932	23.2120	9.9287
1956	90668.0	104.057	25.1838	23.0372	10.0595
1957	91180.0	103.378	24.9862	22.8423	10.1920
1958	89857.0	100.000	25.0246	22.7806	10.3262
1959	92583.0	103.893	25.0292	22.7106	10.4622

(Continued)

TABLE A2—(*Continued*)

	Quality-Adjusted		Years since Migration		
Year	Private Employment (*QAPE*)	Private Hours (*QATHWP*)	Males (Z_M)	Females (Z_F)	Median Education (E)
1960	94413.0	105.538	25.0162	22.6320	10.6000
1961	96286.0	106.387	24.9688	22.5489	10.9927
1962	99939.0	111.287	24.8591	22.4574	11.4000
1963	101860.0	112.876	24.7028	22.3030	11.5499
1964	104460.0	115.742	24.6130	22.1592	11.7000
1965	106803.0	119.463	24.5116	22.0063	11.8000
1966	109417.0	123.261	24.3272	21.8089	12.0000
1967	110564.0	124.772	24.0511	21.5300	12.0000
1968	112961.0	127.142	23.5628	21.0640	12.1000
1969	115492.0	129.834	23.2771	20.8785	12.1499
1970	116095.0	127.015	22.9404	20.6916	12.2000
1971	116022.0	127.206	22.6392	20.5103	12.2000
1972	119019.0	130.250	22.3128	20.3140	12.2000
1973	123084.0	135.403	21.9676	20.0986	12.3000
1974	124682.0	136.016	21.6471	19.9108	12.3000
1975	121653.0	130.721	21.3589	19.7535	12.3000
1976	126290.0	134.864	21.1000	19.8000	12.4000
1977	130682.0	140.578	20.6000	19.4000	12.4000
1978	135750.0	148.357	19.9000	18.9000	12.4000
1979	140376.0	154.452	19.6000	18.7000	12.5000

Sources: *QAPE* was computed using equations (1) through (5) as explained in the text and data from my 1984 study. $QATHWP = (QAPE \cdot AHWP)/QAPE_{1958}$ so that the base year (1958) is 100.0. The Z_M and Z_F were computed using equation (4), 1909 and 1970 benchmarks, and data from my 1984 study. Missing values for E were logarithmically interpolated between the following observations: 1910, 1920, 1930 (John Folger and Charles Nam, 1964, p. 253); 1900, extrapolated by author from above values; 1940, 1950, 1960, *CPR* #356; 1962, 1964, 1965, 1966, *CPR* #158; 1967, *CPR* #169; 1968, *CPR* #182; 1970, *CPR* #207; 1971, *CPR* #229; 1972, *CPR* #243; 1973, 1974, *CPR* #274; 1975–79, *CPR* #356; where *CPR* denotes U.S. Bureau of the Census, *Current Population Reports*, Series P-20, and the issue dates of the reports are #158 December 19, 1966, #169 February 9, 1968, #182 April 28, 1969, #207 November 30, 1970, #229 December 1971, #243 November 1972, #274 December 1974, and #356 August 1980.

REFERENCES

Ackley, Gardner, *Macroeconomic Theory*, New York: Macmillan, 1961.

Barro, Robert J. and Grossman, Herschel I., "Suppressed Inflation and the Supply Multiplier," *Review of Economic Studies*, January 1974, *41*, 87–104.

Blau, Francine, D., "Immigration and Labor Earnings in Early Twentieth Century America," *Research in Population Economics*, 1980, *2*, 21–41.

______, "Discrimination against Women: Theory and Evidence," in William A. Darity, Jr., ed., *Labor Economics: Modern Views*, Boston: Kluwer Nijhoff, 1984.

Chinloy, Peter, "Sources of Quality Change in Labor Input," *American Economic Review*, March 1980, *70*, 108–19.

Chiswick, Barry R., "The Effects of Americanization on the Earnings of Foreign-born Men," *Journal of Political Economy*, October 1978, *86*, 897–921.

______, "The Economic Progress of Immigrants: Some Apparently Universal Patterns," in William Fellner, ed., *Contemporary Economic Problems 1979*, Washington: American Enterprise Institute, 1979.

Christensen, Laurits R. and Jorgenson, Dale W., "U.S. Input, Output, Saving and Wealth, 1929–1977," unpublished paper, Harvard Institute of Economic Research, December 1978.

Cox, Charles, C., "The Enforcement of Public Price Controls," *Journal of Political Economy*, October 1980, *88*, 887–916.

Darby, Michael R., (1976a) "Price and Wage Controls: The First Two Years," and "Further Evidence," in K. Brunner and A. H. Meltzer, eds., *The Effects of Price and Wage Controls*, Rochester-Carnegie Conference Series on Public Policy, Vol. 2, *Journal of Monetary Economics*, Suppl. April 1976, 235–63.

______, (1976b) "The U.S. Economic Stabilization Program of 1971–1974," in M. Walker, ed., *The Illusion of Wage and Price Control*, Vancouver: Fraser Institute, 1976.

______, (1976c) "Three-and-a-Half Million U.S. Employees Have Been Mislaid: Or, an Explanation of Unemployment, 1934–1941," *Journal of Political Economy*, February 1976, *84*, 1–16.

______, *Intermediate Macroeconomics*, New York: McGraw-Hill, 1979.

______, "The Price of Oil and World Inflation and Recession," *American Economic Review*, September 1982, *72*, 738–51.

______, *Labor Force, Employment, and Productivity in Historical Perspective*, Los Angeles: UCLA Institute of Industrial Relations, 1984.

Denison, Edward F., *Accounting for United States Economic Growth, 1929–1969*, Washington: The Brookings Institution, 1974.

______, *Accounting for Slower Economic Growth: The United States in the 1970s*, Washington: The Brookings Institution, 1979.

Evans, Paul, "The Effects of General Price Controls in the U.S. during World War II," *Journal of Political Economy*, October 1982, *90*, 944–66.

Fair, Ray, *The Short-Run Demand for Workers and Hours*, Amsterdam: North-Holland, 1969.

Folger, John K. and Nam, Charles B., "Education Trends from Census Data," *Demography*, 1964, *1*, 247–57.

Friedman, Milton, "A Bias in Current Measures of Economic Growth," *Journal of Political Economy*, March/April 1974, *82*, 431–32.

Gordon, Robert J., "The 'End of Expansion' Phenomenon in Short-Run Productivity Behavior," *Brookings Papers on Economic Activity*, 2:1979, 447–61.

Griliches, Zvi, "Hedonic Price Indexes for Automobiles: An Econometric Analysis of Quality Change," in U.S. Congress, Joint Economic Committee, *Government Price Statistics, Hearings... January 24, 1961*, Washington: USGPO, 1961.

Jwa, Sung Hee, "Towards an Equilibrium Approach to the Effects of Price Controls: A Theoretical and Empirical Analysis of the Effects of Price Controls on Quality Offerings," unpublished doctoral dissertation, University of California-Los Angeles, 1983.

Morris, Charles, S., "Cyclical Productivity in the United States: A Reconciliation of Theory and Evidence," unpublished doctoral dissertation, University of California-Los Angeles, 1983.

Oi, Walter Y., "Labor as a Quasi-Fixed Factor," *Journal of Political Economy*, December 1962, *70*, 538–55.

Perry, George L., "Potential Output and Productivity," *Brookings Papers on Economic Activity*, 1:1977, 11–47.

Rasche, Robert H. and Tatom, John A., "Energy Resources and Potential GNP," *Federal Reserve Bank of St. Louis Review*, June 1977, *59*, 10–24.

______ **and** ______, "Energy Price Shocks, Aggregate Supply and Monetary Policy: The Theory and International Evidence," *Carnegie-Rochester Conference Series on Public Policy*, Spring 1981, *14*, 9–93.

Sims, Christopher A., "Output and Labor Input in Manufacturing," *Brookings Papers on Economic Activity*, 3:1974, 695–728.

Smith, James P. and Welch, Finis R., "Black-White Male Wage Ratios: 1960–70," *American Economic Review*, June 1977, *67*, 323–38.

Solow, Robert M., *Growth Theory: An Exposition*, New York: Oxford University Press, 1970.

______, "Some Evidence on the Short-Run Productivity Puzzle," in J. Bhagwati and R. S. Eckaus, eds., *Development and Planning: Essays in Honor of Paul Rosenstein Rodan*, Cambridge: MIT Press, 1973.

Terborgh, George, "A Quizzical Look at Productivity Statistics," *Capital Goods Review* of the Machine and Allied Products Institute, No. 110, August 1979.

Price Statistics Review Committee of the National Bureau of Economic Research, "The Price Statistics of the Federal Government: Report of the Committee," in U.S. Congress, Joint Economic Committee, *Government Price Statistics, Hearings... January 24, 1961*, Washington: USGPO, 1961.

U.S. Bureau of the Census, *1970 Census of Population: Subject Reports: National Origin and Language*, Report No. PC(2)-1A, Washington: USGPO, 1973.

U.S. Bureau of Economic Analysis, *Long Term Economic Growth, 1860–1970*, Washington: USGPO, 1973.

U.S. Council of Economic Advisers, *Economic Report of the President, 1979; 1980*, Washington: USGPO, 1979, 1980.

U.S. Immigration Commission, *Immigrants in Industries: Part 23: Summary Report on Immigrants in Manufacturing and Mining*, Reports of the Immigration Commission, Vol. 20, Washington: USGPO, 1911.

Product Line Rivalry

By JAMES A. BRANDER AND JONATHAN EATON*

Most firms offer entire product lines rather than single products. There is, however, only a relatively small body of literature devoted to product line selection by multiproduct firms.[1] What literature does exist has focused chiefly on cost considerations: multiproduct firms are seen to emerge principally as the consequence of economies of scope in production. A recent survey of the literature on multiproduct firms (Elizabeth Bailey and Ann Friedlaender, 1982) reflects the recent literature, considering closely only the limited demand effects allowed by the "contestable markets hypothesis." Our view is that interaction between the demands for different products, and the associated strategic effects, are important determinants of the products a single firm will produce.

How is it that product lines are determined? Certainly we see several different patterns in the real world. Probably the most common pattern is that each firm produces a wide range of varieties within a product group, and a number of firms produce very similar, and sometimes virtually identical, products. Ford and General Motors produce closely competing product lines, as do Nikon, Canon, and Minolta in the camera industry, and so on.

Occasionally, however, one firm manages to gain almost exclusive control over a well-defined part of the product spectrum, and does not venture into other parts. An obvious example is Mercedes Benz in the automobile industry, and a trip through a department store will yield a number of other examples. Naturally, more complex versions of this basic pattern arise. A single firm may have several areas of control in a product group, or two or three firms may dominate one part of the product spectrum, while other firms produce less closely substitutable product lines. In addition, a fairly common historical pattern is for firms to expand the scope of their product offerings and compete more directly with each other as the market grows.

Michael Spence (1976), in a paper concerned mainly with single product firms, suggests the result that, in the multiproduct case, close substitutes will be produced by different firms. The reason is fairly straightforward: if firm A produces product 1 and product 2 is a close substitute, then production of product 2 is likely to appear more attractive to firm B than to firm A because B will not be concerned about the consequent reduction in demand for product 1.

However, one wonders, might not firm A recognize that if it doesn't produce product 2, firm B will, and therefore try to preempt B. Strategic preemption requires a two-stage (or more) decision process. In choosing to produce a particular product, firms must anticipate that this will have some effect on later competition. But surely this is precisely the way product selection occurs. As argued by Edward Prescott and Michael Visscher (1977), product selections, once decided upon, are not easily changed. Product selection is a commitment,[2] which, to a first

*Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, B.C., V6T 1Y8, Canada, and Economic Growth Center, Yale University, Box 1987 Yale Station, New Haven, CT 06520, respectively. The motivation for this paper arose from a series of conversations, or more accurately, arguments, with Avinash Dixit and Barbara Spencer. The authors have abandoned many of their original thoughts and now, at least, agree with each other. We are grateful to a referee for helpful comments and suggestions.

[1] The two classic approaches to product selection derive from the work of Harold Hotelling (1929) and Edward Chamberlin (1933). Most of the recent work in these traditions, including the widely cited work of Kelvin Lancaster (1979), Avinash Dixit and Joseph Stiglitz (1977), Michael Spence (1976), and Steven Salop (1979) focuses on the one product per firm case.

[2] The term commitment (or "credible threat") refers to the idea that in strategic interaction, a firm (or player) might reasonably be expected to believe that a rival will only pursue actions that are in the rival's best interests. Threats that can only be carried out through suboptimal behavior (as in the Sylos-Labini limit output model) are

approximation, is taken as given in the following output or price rivalry.

In this paper we make a series of straightforward but significant points about product selection by multiproduct firms. In particular, using a very simple structure, we find that sequential decisions on product type and output can naturally give rise to equilibria in which a single firm monopolizes close substitutes. Such outcomes hold only for certain levels of demand and might, therefore, be observed only over some portion of the life cycle of the industry.

Perhaps the most closely related existing work is in the literature on preemption in product space, particularly (in addition to Prescott and Visscher) Donald Hay (1976) Curtis Eaton and Richard Lipsey (1979) and W. J. Lane (1980). Eaton and Lipsey (1979) have the idea that a foresighted monopolist would introduce a near product in a growing market before a rival. Our paper is in a similar spirit, but is concerned principally with cases in which firms have equal opportunity.

Section I sets out the basic model, Section II briefly considers monopoly, Section III derives the main results on product line rivalry, and Section IV presents a result on product line choice and entry deterrence.

I. The Model

To examine product line rivalry, we consider a constellation of four possible products. Two products are close substitutes for each other and more distant substitutes for the other pair, which are, in turn, close substitutes for each other. In particular, commodity pairs (1,2) and (3,4) are close substitutes, while pairs (1,3), (1,4), (2,3), and (2,4) are more distant substitutes. This is about the simplest structure in which the question of whether competing multiproduct firms produce close or distant substitutes can be addressed.

We use inverse demand functions and define closeness of substitutes using the cross derivatives of these functions. The price of good i is denoted p^i, the quantity of good i is denoted x^i and X is the vector (x^1, x^2, x^3, x^4). The inverse demand function is then written

$$(1) \qquad p^i = p^i(x^1, x^2, x^3, x^4) = p^i(X).$$

To focus as clearly as possible on the essential issues, we assume that the demand structure is perfectly symmetric except for the differences in substitutability already described. In saying that goods 1 and 2 are closer substitutes[3] than 1 and 3, we mean that the response of p^1 to a change in the output of good 2 is greater in absolute value than the response to a change in x^3. Since these goods are substitutes, the cross-price effects are negative, and, using subscripts to denote derivatives, we have

$$(2) \qquad p^i_j < p^i_k,$$

where i and j are close substitutes and i and k are more distant substitutes. We could imagine that the degree of substitutability might vary with X and that goods that were close substitutes in some ranges might be

not credible. This idea goes back at least as far as Thomas Schelling (1956), but has only received attention recently. Recent work includes Spence (1977, 1979), James Friedman (1979), Dixit (1980) and B. Curtis Eaton and Richard Lipsey (1979, 1980, 1981).

[3] It is perhaps more normal to express the substitutability relationship using the ordinary demand functions $x^i(p)$, in either derivative or elasticity form. As pointed out by John Hicks (1956, pp. 156 ff.), the two definitions do not necessarily coincide, and neither definition is more "natural" than the other. Hicks referred to substitutes by our definition as "q-substitutes" and by the other definition as "p-substitutes." The (compensated) substitution effects are also sometimes referred to as "Antonelli substitution effects" and "Slutsky substitution effects," respectively, because they are elements of the Antonelli matrix and the Slutsky matrix, respectively. The Antonelli and Slutsky matrices are generalized inverses of each other. (See Angus Deaton and John Muellbauer, 1980, p. 57.) Since we are going to assume that firms use quantity rather than price as a choice variable, the q-substitute approach is more convenient. One could, however, use prices as the choice variables and use p-substitutes.

We use the derivative rather than the elasticity form for the examination of substitution effects because it is the derivative that appears directly in the mean value theorem and in the first-order conditions that are used in the analysis.

distant substitutes in others. To make the questions we wish to address well-defined without complication, (2) is assumed to hold uniformly: p^i_j evaluated anywhere in the range of interest is greater in absolute value (more negative) than p^i_k evaluated anywhere in the range of interest.

One example of a computationally simple demand structure with the properties described here arises from utility that is quadratic in X and additively separable from a numeraire good: $u = aX - X^TBX + m$, where b_{12} and b_{34} are equal and exceed other off-diagonal elements of the symmetric matrix B, which are themselves equal, and m is consumption of a numeraire good. We use this functional form in Section IV.

A firm that produces any of the four products must have made three decisions: (*i*) how many products to produce (the scope decision); (*ii*) which particular products to produce (the line decision); and (*iii*) which quantity to produce (or which price to charge) for each product. As logical possibilities we might imagine that these decisions could be made sequentially, or that the first, or second two, or even all three, could be made simultaneously. Which assumption is appropriate in any particular case depends on actual technological considerations. In our analysis, we treat the product line decision as strictly prior to the final price or quantity choice. In other words, firms establish prices or quantities taking their own and their rivals' line and scope decisions as given.

The final stage may, as indicated, be either a price decision or a quantity decision.[4] For concreteness, in this paper we take quantity as the third-stage decision variable. We have examined a number of our results when price is the final decision variable, and the central insights of our analysis were not affected, although there are some relatively minor differences which we mention at later points in the paper.

The overall equilibrium concept we work with for the most part is the subgame perfect equilibrium. This equilibrium concept incorporates two important ideas: first, the equilibrium is noncooperative so that at each stage equilibrium occurs when each firm is maximizing profit, given the current and previous strategy variable levels chosen by its rival, and second, each firm understands at any stage how future stages will be affected by current decisions.[5]

Our assumptions about technology are very simple. Each firm incurs a sunk cost K for each product it plans to produce at the time of the scope decision. (Indeed, it may be this sunk cost which contributes to making the scope decision credible.) A constant marginal cost c is incurred at the time quantity decisions are made. We assume K and c to be the same for all four products. Note that, while there is interaction among demands for the four commodities, we assume that their cost structures are independent: in particular, there are no economies or diseconomies of scope.

II. Monopoly

Although we are principally concerned with the rivalry between firms, there is one important insight to be established for the monopoly case. Specifically, if a monopolist chooses to produce only two products, it will choose two distant substitutes rather than two close substitutes. The reasoning involved is fairly straightforward, but it is worth being precise. Imagine that the monopolist is producing products 1 and 2, which are close substitutes, at the profit-maximizing levels. By symmetry, $x^1 = x^2 (= x)$. Holding the output level of product 1 fixed, imagine re-

[4] Whether price or quantity Nash equilibria are appropriate depends on the nature of production. Indeed, it may be useful to think of quantity and price as occurring sequentially. If a quantity decision is a credible commitment, due perhaps to practical irreversibilities in production and high storage costs, and price later clears the market, then quantity should be regarded as the third-stage decision variable. Alternatively, if a price announcement is a credible threat, with quantity being the residual variable that clears the market, the third stage should be modeled as a price game. (This interpretation is in Friedman, 1980.) For some other comments on the issue of price vs. quantity Nash equilibria, see Timothy Bresnahan (1981).

[5] The concept of subgame perfection is associated with R. Selten (1975) and has been the focus of considerable recent attention in the oligopoly literature. A good pedagogical discussion is in Martin Shubik (1982).

placing production of x^2 with an equal amount of x^3. Let $X' = (x, x, 0, 0)$ and let $X'' = (x, 0, x, 0)$. The effect on the price of good 1 is as follows:

$$(3) \qquad \Delta p^1 = p^1(X'') - p^1(X').$$

Using the mean value theorem,[6] this difference can be expressed as

$$(4) \qquad \Delta p^1 = \left(p^1_3(X^*) - p^1_2(X^*)\right)x$$

for some $X^* \in [X', X'']$. From (2), the price change must be positive and, by symmetry, the other price must also rise. Consequently, even without optimal readjustment of quantities, prices and profits must rise. A monopolist who plans to produce just two goods, and who is unconcerned about entry, will therefore produce two distant substitutes. This result serves as a useful base for comparison with the two-firm case.

III. Three-Stage Duopoly

A. *The Output Decision*

We now discuss the duopoly equilibrium when firms make the scope, line, and output decisions in sequence. We examine the decisions in reverse order, starting with the output decisions of firms already committed to particular line and scope decisions. The third stage is modeled as a Nash quantity (or Cournot) game, taking line and scope decisions of both firms as given. There are many possible configurations of scope and line with which firms might enter the quantity stage. Each of the four products might be produced by firm *A*, by firm *B*, by both firms, or by neither firm, giving rise to 256 possibilities. Many configurations are, however, isomorphic to one another, and many are relatively uninteresting in that they do not bear on the questions of interest, as, for example, when one firm produces all four products and the other nothing.

The most interesting situations for our analysis are those in which each firm produces two products. Firm *A* might produce one pair of close substitutes while firm *B* produced the other pair. We refer to this case as market segmentation. An alternative, market interlacing, occurs when each firm produces two less closely related products, as, for example, if firm *A* produces goods 1 and 3 while firm *B* produces goods 2 and 4. Note that segmentation or interlacing is determined in the line stage, and is taken as given when final output levels are being determined.

Consider the firms' profits in the two cases. Firm *A*'s profit under segmentation is

$$(5) \quad \pi^s = p^1x^1 + p^2x^2 - c(x^1 + x^2) - 2K,$$

where, for concreteness, firm *A* is assumed to produce goods 1 and 2. The superscript *s* denotes segmentation. The first-order condition associated with product 1 can be written

$$(6) \qquad \pi^s_1 = MR^1 + x^2p^2_1 - c = 0,$$

where $MR^1 = x^1p^1_1 + p^1$, is own-marginal revenue. Second-order conditions are

$$(7) \qquad \pi^s_{ii} < 0, \qquad i = 1,2;$$

$$(8) \qquad \pi^s_{11}\pi^s_{22} - \pi^s_{12}\pi^s_{21} > 0.$$

The first-order condition for product 2 is similar to (6). Firm *B* (producing products 3 and 4) has symmetric first- and second-order conditions. Each first-order condition shows, implicitly, the profit-maximizing choice for one product, given the output levels of the others. The solution of these four reaction functions is the (noncooperative) Nash equilibrium in quantities. We assume that demand is sufficiently regular that there exists a unique equilibrium. This equilibrium must, then, be symmetric: all quantities are equal (as are all prices). We assume also that own-marginal revenue declines when the output of any other good rises:

$$(9) \qquad MR^i_j < 0.$$

[6] This is the "mean value theorem for several variables" as described by, for example, Marcel Rosenlicht (1968). For a similar application of this theorem in economics, see Barbara Spencer (1979).

This is a natural condition which holds for most (but not all) plausible demand structures.

Under interlacing, firm A produces, let us say, products 1 and 3. (One can substitute i and k to achieve generality.) This leads to the first-order condition

$$\pi_1^t = MR^1 + x^3 p_1^3 - c = 0, \tag{10}$$

and to similar second-order conditions as before. The superscript t denotes interlacing. The fundamental comparative property of segmentation and interlacing is expressed in Proposition 1.

PROPOSITION 1: *The segmented structure gives rise to higher prices and profits than the interlaced structure.*

PROOF:

By symmetry all products sell for the same price, denoted p^s under segmentation and p^t under interlacing. There are three possibilities: $p^t = p^s$, $p^t > p^s$, or $p^t < p^s$. The first two lead to contradictions. Consider first the equality case. If prices are equal, then quantities must also be equal in the two regimes, and so must MR^1. However, $p_1^2 < p_1^3$ so (6) and (10) cannot both be satisfied. Therefore p^t cannot equal p^s.

Now consider $p^t > p^s$. This implies $x^t < x^s$ and, by (9), that MR^1 is larger under interlacing. Since $p_1^2 < p_1^3 < 0$, it follows once again that (6) and (10) could not both be satisfied. Therefore $p^t < p^s$ as was to be shown.

Since $p^t < p^s$, it follows that $x^t > x^s$. Consider now total profits

$$\pi = \sum p^i x^i - c\left(\sum x^i\right) - 4K,$$

and, for concreteness, but without loss of generality, consider a change in π due to an increase in x^1:

$$\partial \pi / \partial x^1 = p^1 + \sum x^i p_1^i - c.$$

At the segmented solution $\partial \pi / \partial x^1 = x^3 p_1^3 + x^4 p_1^4 < 0$ since $MR^1 + x^2 p_1^2 - c = 0$ from (6). Similarly, $\partial \pi / \partial x^1$ is also negative at the interlaced solution using (10) rather than (6). Clearly π is decreasing in all xs, not just in x^1, between x^t and x^s. It then follows, given overall concavity of π in the x^is, that profit falls as outputs increase in moving from the segmented to the interlaced solution.

These interlaced and segmented structures are only two of many possibilities. Even confining attention to the scope structure of two products per firm, we might imagine that the same product or products could be produced by both firms, leaving one or two products unproduced. The profit of at least one firm in such cases would normally fall short of its profit even in the interlaced case, given the symmetric structure of demand. There is also a series of cases in which each product is produced by at least one firm, with some overlapping in the sense that some products are produced by both firms. We defer consideration of these and other cases for the present, and move on to consideration of the line decision.

B. *The Line Decision*

When making the line decision, firms take the scope decision of how many products to produce as fixed, and have only to decide upon which products to produce. The case for separation of the line and scope decisions is not as compelling as that for separation of line and output decisions. Nevertheless, it seems to capture an important flavor of real product selection in that firms often commit themselves to a particular market, especially to new or anticipated markets, well before actual product types are decided upon. Separation of line and scope is not crucial to the analysis in any case, but it is our feeling that the full three-stage model brings out the logical structure of the argument most clearly.

When making the product line decision, firms are aware that they will be involved in a noncooperative output game in the future and take this into account in the line stage. We examine the case in which each firm is committed, from the scope stage, to producing two products. Consider the Nash equilibrium in the game in which firms choose product lines simultaneously. For a wide range of demand structures (although not all), the Nash reaction to a rival's two products is simply the other two products. We restrict our attention first to this case, for which it follows that any 2×2 division is a Nash equilibrium, and, in particular, that market segmentation and market interlacing are equilibria.

PROPOSITION 2: *Both market segmentation and market interlacing are Nash equilibria in the simultaneous product selection game.*

The reason that both segmented and interlaced structures are observed may simply be that both are Nash equilibria at the line stage and can therefore be part of a subgame perfect equilibrium structure for the entire game. Any 2×2 division has a certain inertia associated with it. This simple insight itself seems a worthwhile addition to the Spence (1976) discussion. When the natural sequencing of product selection and output rivalry are taken into account, it is quite possible that close substitutes will be produced by a single firm.

We now consider two modifications of the model presented so far, both of which strengthen the case for segmentation. The first modification involves a change in the game being played. Specifically, the line decision process is assumed to be asymmetric: one firm is able to choose its two products first. This may occur because of random factors in the product selection process, or for some other exogenous reason. Formally the game becomes a four-stage game. As part of the subgame perfect equilibrium structure, the first firm to choose knows that its rival will act in its own best interest in the next stage. If it knows that whichever products it chooses, its rival will choose the other two,[7] then the first firm is in a position to choose either market segmentation or market interlacing as the industry structure. Since market segmentation leads to higher profits for each firm, Proposition 3 is immediate.

PROPOSITION 3: *If firms enter the line stage sequentially rather than simultaneously, market segmentation is the Nash equilibrium.*

The second modification we consider is a change in the equilibrium concept. We return to the game in which product line choices are made simultaneously, but instead of examining the Nash equilibrium, we examine a Stackelberg leader-leader equilibrium.

We define a Stackelberg strategy as one which involves taking into account the contemporaneous reaction of one's rival in setting one's own strategy.[8] In the output case, if both players follow Stackelberg strategies, the outcome is not an equilibrium because both firms choose an output other than what the other expects. However, if a firm can correctly assume that if it chooses two products, its rival will choose the other two, then a Stackelberg strategy at the line stage leads it to choose two close substitutes. The other firm, also following a Stackelberg strategy, will be doing the same thing. The joint Stackelberg equilibrium arises when the firms choose different pairs.

PROPOSITION 4: *The product line game has a joint Stackelberg equilibrium, and this equilibrium is characterized by market segmentation.*

There are two interpretations one might place on this Stackelberg equilibrium. On one hand, we can imagine that each firm literally expects the other to be a Nash follower. Firms therefore have an incorrect view of their rivals' behavior, but "by accident" select consistent actions. In equilibrium, firms are "right for the wrong reasons." This inter-

[7] This market structure is similar to the one examined by Prescott and Visscher, who considered sequential entry by firms precommitted to a single product. Here we consider sequential entry by firms precommitted to two products. Prescott and Visscher also assume that the entry and line decisions are simultaneous rather than sequential as we assume here.

[8] The term Stackelberg is sometimes used to mean that one player acts before another. Stackelberg's original model can be interpreted in either way, and usage seems to be divided. The product line game itself is an example of what is sometimes called a "game of coordination." See Shubik.

pretation then has the same appeal but also the same weakness as the traditional interpretation of the Nash equilibrium: out of equilibrium, firms' expectations about their rivals would be false.

An alternative interpretation is that firms understand the incentive structure in which they operate and recognize that other firms are much like themselves. Each firm knows that the other is trying to jockey it into selecting two closely related products. However, this coincides with each firm's own objectives and therefore leads to a consistent equilibrium in which firms are "right for the right reasons." In essence, firms accept that only Nash equilibria are individually rational and therefore cannot expect to achieve the collusive outcome. However, when there are two Nash equilibria, they can select the Nash equilibrium that strictly dominates the other for both. Each firm is a Stackelberg leader, not in the sense that it attributes false reactions to its rival, but in the sense that it correctly anticipates what the other is trying to do and is able to act on this knowledge.[9] This is the nature of the joint Stackelberg solution, and leads to a strong presumption in favor of the segmented solution, given the initial scope decision of two products each, and the assumption that the best response to any product pair selection is the other pair.

The joint Stackelberg solution, because it coincides with a Nash equilibrium, is self-enforcing and therefore credible in earlier stages, and as a result is admissible as part of a perfect foresight structure.

As we have emphasized, Propositions 2, 3, and 4 assume that the "Nash reaction" at the line stage to a rival's product choice is the remaining two products. If rivalry at the final stage is Bertrand (Nash competition in prices), this is always so. If the final stage is, as we assume, Cournot (Nash in outputs), then the Nash response to any two products is not always the other two. For example, if products 1 and 2 are very close substitutes for each other, but virtually unrelated to products 3 and 4, then if firm A chooses products 1 and 2, firm B would choose products 1 (or 2) and 3 (or 4). In the limit, as 1 and 2 became perfect substitutes that were completely unrelated to products 3 and 4, which were perfect substitutes for each other, then firm B would have nothing to gain by producing product 3, given that it planned to produce product 4, so it would prefer to overlap firm A's product set.[10]

For such examples, our Propositions 2, 3, and 4 no longer obtain. Market segmentation is not a Nash equilibrium, while interlacing is not only the unique Nash equilibrium but the sequential entry equilibrium and joint Stackelberg equilibrium as well. Firm A, knowing that if it chooses products 1 and 2, firm B will choose 1 or 2 as well, will prefer to choose 1 and 3, to which the Nash response of firm B is product 2 and 4: interlacing emerges.

What is required to preclude such examples? The Nash response to a product pair is always the other pair as long as the "close" products are reasonably differentiated from each other, and not too strongly differentiated from the other pair. For example, using the quadratic utility function introduced in Section II, if $a=5$, $b_{ii}=1$ $(i=1,2,3,4)$ and $b_{12}=b_{34}=.4$, the best response to a choice of products 1 and 2 would be 3 and 4 as long as the elements $b_{13}=b_{14}=b_{23}=b_{24}$ exceed .05.

[9] At this point, some comment on terminology is perhaps appropriate. Instead of referring to this equilibrium concept as a joint Stackelberg equilibrium, one might be tempted to use the term "rational expectations" equilibrium or "consistent conjectures" equilibrium. If by rational expectations, one means that "the players know the model," then this term would be appropriate. Unfortunately the term rational expectations has been used in a number of different senses. We prefer, in microeconomics at any rate, that the term "rational expectations equilibrium" refer to a situation in which players' expectations about the levels of strategy variables chosen by rivals are confirmed in equilibrium. Thus any Nash equilibrium would be a rational expectations equilibrium. The consistent conjectures term (compare Bresnahan, Martin Perry, 1982, and Morton Kamien and Nancy Schwartz, 1983, among others) captures the idea that firms should have correct anticipations about levels of rivals' strategy variables in equilibrium and about responses of these strategy variables to change in one's own strategy variable. We prefer the term "joint Stackelberg" for our particular model.

[10] We are grateful to Simon Anderson for pointing out this example to us.

C. *The Scope Decision*

Why should the firms settle on two products each? Consider the Nash reaction functions for the scope decision. Demand may be sufficiently low that if one firm commits to only one product, the other firm would prefer not to enter at all. On the other hand, demand may be so great that even if one firm committed itself to all four products, the optimal response of its rival would be to produce all four products also: complete overlapping.[11] Only if demand is in that intermediate range where the optimal response to a scope decision of two is also two can the equilibrium structure described in the previous subsection emerge.

A subgame perfect equilibrium is a Nash equilibrium in the overall (three-stage) game that is also a Nash equilibrium in any subgame; that is, in the two-stage game of product line and output choices and in the one-stage game of output choice alone. Therefore, the point remains that there are ranges of demand for which market segmentation is a subgame perfect equilibrium. Similarly, the subgame perfect equilibrium for the four-stage game involving sequential entry in the line stage is characterized by market segmentation for some demand levels, as is the equilibrium incorporating a joint Stackelberg solution in the product line stage. However, as growth occurred in the market and the game were repeated, market segmentation would be replaced by market overlapping.

We have presented a simple model which we believe throws some light on product line rivalry between firms. We find that market segmentation is a very reasonable outcome once the multistage structure of market rivalry is explicitly recognized, although many other configurations are possible, including overlapping of firms. In the next section we discuss how the threat of further entry can increase the likelihood that interlacing rather than segmentation is the outcome.

IV. Competition as Entry Deterrence

Our analysis thus far has assumed that at the time firms make their line decisions, there is no possibility of further entry. Relaxing this assumption increases the likelihood that the outcome of sequential product line choice or a joint Stackelberg equilibrium is one of market interlacing: firms that have already entered and made a scope decision may deliberately choose an interlaced structure to make the market more competitive, reducing the profitability of further entry.

Consider again the constellation of four products of equation (1) and the product line decisions of two firms, *A* and *B*, each having a scope of two products. Propositions 3 and 4 established the presumption in favor of segmentation in the absence of a threat of entry. Assume, however, that further entry is possible after firms *A* and *B* have made their line decisions. Consider, for simplicity, the case of a single firm entering and establishing production of just one of the four products. Because of the symmetry of our specification, it does not matter which one, so assume that it is product 1. If firm *A* has committed itself to products 1 and 2 and firm *B* to products 3 and 4 (the segmented case), then the profits of the three firms will be given by

$$(11)\quad \pi^{ASE} = p^1(X)x_A^1 + p^2(X)x^2 - c\left(x_A^1 + x^2\right) - 2K,$$

$$(12)\quad \pi^{BSE} = p^3(X)x^3 + p^4(X)x^4 - c(x^3 + x^4) - 2K,$$

$$(13)\quad \pi^{ESE} = p^1(X)x_E^1 - cx_E^1 - K,$$

where

$$(14)\qquad X = \left(x_A^1 + x_E^1, x^2, x^3, x^4\right),$$

and the outputs are at their Cournot equilibrium values. Here x_A^1 denotes the output of commodity 1 produced by firm *A* and x_E^1

[11] This last result depends upon our assumption that competition at the final stage is a Nash game in outputs (a Cournot-Nash game). Were it a Nash game in prices (a Bertrand-Nash game), then no more than one firm would ever produce the same product.

that produced by the entrant; π^{ASE}, π^{BSE}, and π^{ESE} denote equilibrium profits of firm A, firm B, and the entrant, respectively, under market segmentation with entry.

Under market interlacing, with firm A committed to producing products 1 and 3 and firm B to 2 and 4, after-entry profits will be given by

$$(15)\quad \pi^{AIE} = p^1(X)x_A^1 + p^3(X)x^3 - c\left(x_A^1 + x^3\right) - 2K,$$

$$(16)\quad \pi^{BIE} = p^2(X)x^2 + p^4(X)x^4 - c\left(x^2 + x^4\right) - 2K,$$

$$(17)\quad \pi^{EIE} = p^1(X)x_E^1 - cx_E^1 - K,$$

where now π^{AIE}, π^{BIE}, and π^{EIE} denote the equilibrium levels of the three firms' profits when product lines are interlaced. The value X continues to be defined by (14), and the outputs assume their Cournot values under interlacing.

Finally, under market interlacing without entry, firms A and B will earn π^{AI} and π^{BI} given by

$$(18)\quad \pi^{AI} = p^1(X)x^1 + p^3(X)x^3 - c\left(x^1 + x^3\right) - 2K,$$

$$(19)\quad \pi^{BI} = p^2(X)x^2 + p^4(X)x^4 - c\left(x^2 + x^4\right) - 2K,$$

$$(20)\quad X = \left(x^1, x^2, x^3, x^4\right).$$

We now state:

PROPOSITION 5: *With the threat of further entry, the interlaced structure can give rise to higher prices and profits for both incumbents than the segmented structure.*

PROOF:

This results obtains if: 1) under segmentation entry is profitable ($\pi^{ESE} > 0$), 2) under interlacing it is not ($\pi^{EIE} < 0$), and 3) firms A and B earn higher profits with an interlaced structure and no entry than with a segmented structure with entry ($\pi^{AI} > \pi^{ASE}$ and $\pi^{BI} > \pi^{BSE}$). To establish that these three conditions can be satisfied simultaneously, we present an example using the linear demand structure mentioned in Section I: $u = aX - X^TBX + m$. (The calculations are long and tedious, and were done on a computer.) When $b_{ii} = 1$, $i = 1,2,3,4$; $b_{12} = b_{34} = .3$ $b_{13} = b_{14} = b_{23} = b_{24} = .1$, $a = 5$, $c = 2$, and $K = .4$, we obtain, under segmentation with entry: $P = (2.93, 3.26, 3.37, 3.37)$; $\pi^{ASE} = 0.16$; $\pi^{BSE} = 0.64$; $\pi^{ESE} = 0.03$. Under interlacing with entry: $P = (2.85, 3.20, 3.23, 3.25)$; $\pi^{AIE} = 0.23$; $\pi^{BIE} = 0.56$; $\pi^{EIE} = -0.04$. Finally, under interlacing without entry: $P = (3.27, 3.27, 3.27, 3.27)$; $\pi^{AI} = 0.66$; $\pi^{BI} = 0.66$.

At these values, market segmentation permits entry while interlacing does not. The initial two entrants earn higher profits under interlacing without entry than under a segmented structure with entry.

Perhaps it is not surprising that firm A earns a higher profit under interlacing without entry than under segmentation with entry, since it is the firm that ends up sharing a product with the entrant. More surprising is that even firm B can earn a higher profit under the first configuration than under the second.

What conditions are likely to lead to market interlacing as entry deterrence? One is that the degree of substitutability between the two pairs of "close" products substantially exceed that between other pairs. Otherwise the attractiveness of entry by a third firm is unlikely to be affected by the first two firms' decision to interlace or segment. If entry is unprofitable under either configuration, segmentation is the perferred equilibrium structure for firms A and B.

As long as both firms A and B experience lower profits under interlacing without entry than under segmentation with entry, and as long as interlacing does effectively deter entry, then interlacing will emerge as the sequential entry equilibrium and the joint Stackelberg equilibrium. There are, of course, cases in which entry-deterring interlacing would be preferred by firm A but not by firm B, provided both firms knew that one of firm A's products would be the target for the entrant. However, as suggested by the doctrine of insufficient reason, firms A and B might reasonably regard their products as equally likely targets for a later entrant. This reinforces the symmetric preference for interlacing. Furthermore, any risk aversion on the part of firms A and B will tend to make interlacing more likely as a sequential entry equilibrium or a joint Stackelberg equilibrium.

It is important to note that we are considering situations in which a third firm commits itself to entry after firms A and B have made their product line decisions. If the entrant had committed itself beforehand, a threat by firms A and B to interlace is not credible. The solution will, in our example, again be one of segmentation.

V. Concluding Remarks

This paper has focused on the (in our view) much neglected subject of product line selection by multiproduct firms. We have restricted attention to "demand-side" influences on product selection. It is fairly clear that "cost-side" considerations are also very important. In this paper, products are independent on the cost side, but if there were, for example, economies of scope between particular products, there would clearly be a stronger incentive for one firm to produce these products. Oil refineries produce a spectrum of different fuels, from heavy oil to light fuels like kerosene, because they are all byproducts of each other: a fairly strong form of economies of scope.

There is a substantial recent literature on economies of scope and multiproduct firms culminating in the 1982 book by William Baumol, John Panzar, and Robert Willig (B-P-W). One other important difference between their work and the present paper, aside from the role of the cost side in the analysis, is the assumption concerning the expectations of firms. In B-P-W, before a firm enters, it takes the current price of each product as given; it is as if scope, line, and price decisions were all made simultaneously. Not surprisingly, this assumption yields an outcome with some resemblance to perfect competition. Our assumption is rather different: firms understand, before anything is actually produced, how the noncooperative output game will work out.

Our basic message is that recognizing the sequential nature of decision making is important in understanding product line rivalry. Market segmentation, in which each firm controls a certain part of the product spectrum, is an equilibrium outcome, although it will only be observed over some fraction of the life cycle of the industry. An interesting extension suggests itself if we consider the possibility of further entry beyond the first two firms in an industry. An interlaced structure, in which close substitutes are produced by different firms, is a more competitive structure than segmentation. More to the point, it is a commitment to greater competition from the point of view of an additional potential entrant. The entrant might therefore be deterred from entry in an interlaced market when it would enter a segmented market: competition as entry deterrence.

Our results have implications for the research and development activity of firms that is aimed at introducing new products. Our theory suggests that a firm that is guaranteed a monopoly over a range of potential products will (provided that demand is uniform) seek to develop those products that are most distant substitutes for what it is currently producing. Production of these products will reduce demand for the monopolist's current products least. When production of a range of potential products is limited to a group of competing firms that are established in the market, each firm is likely to seek to develop products that are close substitutes for what it currently produces, since joint production of these products will lead to less intense price and output competition at a later stage. Finally, if there is threat of entry by firms currently outside the market, each firm may seek to develop products that are more distant substitutes *because* the consequent competition may be so intense as to deter entry.

The analysis of this paper is based on a rather specific formulation, and therefore the results should be interpreted with caution. Generalization to more complex product sets and cost structures would complicate the analysis considerably. Nevertheless, we feel that the extended example developed here identifies central economic tendencies which would continue to operate in more general settings.

REFERENCES

Bailey, Elizabeth E. and Friedlaender, Ann F., "Market Structure and Multiproduct Industries," *Journal of Economic Literature*, September 1982, *20*, 1024–48.

Baumol, William, Panzar, John and Willig, Robert, *Contestable Markets and the Theory of Industry Structure*, San Diego: Harcourt Brace Jovanovich, 1982.

Brander, James A. and Spencer, Barbara J., "Strategic Commitment with *R&D*: The Symmetric Case," *Bell Journal of Economics*, Spring 1983, *13*, 225–35.

Bresnahan, Timothy F., "Duopoly Models with Consistent Conjecture," *American Economic Review*, December 1981, *71*, 934–45.

Chamberlin, Edward H., *The Theory of Monopolistic Competition*, Cambridge: Harvard University Press, 1933.

Deaton, Angus and Muellbauer, John, *Economics and Consumer Behaviour*, Cambridge: Cambridge University Press, 1980.

Dixit, Avinash, "The Role of Investment in Entry-Deterence," *Economic Journal*, March 1980, *90*, 95–106.

______ **and Stiglitz, Joseph,** "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, June 1977, *67*, 297–308.

Eaton, B. Curtis and Lipsey, Richard, G., "The Theory of Market Preemption: Barriers to Entry in a Growing Spatial Market," *Economica*, May 1979, *46*, 149–58.

______ **and** ______, "Exit Barriers are Entry Barriers: The Durability of Capital as a Barrier to Entry," *Bell Journal of Economics*, Autumn 1980, *10*, 721–29.

______ **and** ______, "Capital Commitment and Entry Equilibrium," *Bell Journal of Economics*, Autumn 1981, *11*, 593–604.

Friedman, James W., *Oligopoly and the Theory of Games*, Amsterdam: North-Holland, 1977.

______, "On Entry-Preventing Behaviour and Limit Price Models of Entry," in S. Brams et al., eds. *Applied Game Theory*, Vienna: Springer, 1979.

______, "Advertising and Oligopolistic Equilibrium," mimeo., University of Rochester, 1980.

Hay, Donald A., "Sequential Entry and Entry-Deterring Strategies in Spatial Competition," *Oxford Economic Papers*, July 1976, *28*, 240–57.

Hicks, John A., *A Revision of Demand Theory*, London: Oxford University Press, 1956.

Hotelling, Harold, "Stability in Competition," *Economic Journal*, March 1929, *34*, 41–57.

Kamien, Morton and Schwartz, Nancy, "Conjectural Variations," *Canadian Journal of Economics*, May 1983, *16*, 191–211.

Lancaster, Kelvin, *Variety, Equity and Efficiency*, New York: Columbia University Press, 1979.

Lane, W. J., "Product Differentiation in a Market with Endogenous Sequential Entry," *Bell Journal of Economics*, Spring 1980, *11*, 237–60.

Perry, Martin K., "Oligopoly and Consistent Conjectural Variations," *Bell Journal of Economics*, Spring 1982, *13*, 197–205.

Prescott, Edward C. and Visscher, Michael, "Sequential Location Among Firms with Foresight," *Bell Journal of Economics*, Autumn 1977, *8*, 378–93.

Rosenlicht, Maxwell, *Introduction to Analysis*, Glenview: Scott, Foresman and Company, 1968.

Salop, Steven C., "Monopolistic Competition with Outside Goods," *Bell Journal of Economics*, Spring, 1979, *10*, 141–56.

Schelling, Thomas, "An Essay on Bargaining," *American Economic Review*, June 1956, *46*, 557–83.

Schmalensee, Richard, "Entry Deterrence in the Ready-to-Eat Breakfast Cereal Industry," *Bell Journal of Economics*, Spring 1978, *9*, 305–27.

Selten, R., "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 1975, *4*, 25–55.

Shaked, Avner and Sutton, John, "Relaxing Price Competition Through Product Differentiation," *Review of Economic Studies*, January

1982, *49*, 3–13.

Shubik, Martin, *Game Theory in the Social Sciences: Concepts and Solutions*, Cambridge: MIT Press, 1982.

Spence, A. Michael, "Product Selection, Fixed Costs, and Monopolistic Competition," *Review of Economic Studies*, June 1976, *43*, 217–35.

______, "Entry, Investment and Oligopolistic Pricing," *Bell Journal of Economics*, Autumn 1977, *8*, 534–44.

______, "Investment, Strategy and Growth in a New Market," *Bell Journal of Economics*, Spring 1979, *10*, 1–19.

Spencer, Barbara J., "Asymmetric Information and Government Bureaucracy," unpublished doctoral dissertation, Carnegie-Mellon University, 1979.

Risk, Inflation, and the Stock Market

By ROBERT S. PINDYCK*

From January 1965 to December 1981, the New York Stock Exchange Index declined by about 68 percent in real terms. Including dividends, the average real return as measured by this index was close to zero. Most explanations of this performance focus on the concurrent increase in the average rate of inflation.[1] For example, Franco Modigliani and Richard Cohn (1979) suggested that investors systematically confuse real and nominal discount rates when valuing equity,[2] Eugene Fama (1981) associates higher inflation rates with changes in real variables that reduce the return on capital, and Martin Feldstein (1980a) argued that increased inflation reduces share prices because of the interaction of inflation with the tax system.

Feldstein (1980a, b) and Lawrence Summers (1981a, b) claim that this last effect can explain a large fraction of the decline in share prices. The main sources of the effect are the "historic cost" method of depreciation and the taxation of nominal capital gains, both of which cause the net return from stocks to fall when inflation rises. However, inflation also reduces the real value of the firm's debt, and reduces the net real return on bonds. The size and direction of the overall effect has been debated, and it clearly depends on the values of tax and other parameters.[3] I will argue that increases in expected inflation—together with concurrent increases in the variance of inflation—should have had a small and possibly positive effect on share values.

Burton Malkiel (1979) suggested another reason for the decline in share prices: changes occurred in the U.S. economy during the 1970's that substantially increased the riskiness of capital investments.[4] This paper elaborates on and supports Malkiel's suggestion. It argues that the variance of the firm's real gross marginal return on capital has increased significantly since 1965, that this has increased the relative riskiness of investors' net real returns from holding stocks, and that this in turn can explain a large fraction of the market decline.

The volatility of stock returns has indeed increased, as illustrated by Figure 1, which shows the variance of the total monthly nominal return on the New York Stock Exchange Index, exponentially smoothed around a linear trend line ($\tilde{\sigma}_t^2 = .1\sigma_t^2 + .9\tilde{\sigma}_{t-1}^2$), for 1950–81. (The computation of the sample variance is discussed in Section II.) Also shown is the linear trend line $\bar{\sigma}_t^2 = .000764 + 4.369 \times 10^{-6}t$, which was fitted to the unsmoothed data. Observe that the variance has fluctuated widely, but has roughly doubled over the past twenty years. If shares are rationally valued, this reflects an increase in the variance of firms' gross marginal return on capital, and/or an increase in the variance of inflation.

*Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139. This research was supported by the National Science Foundation under grant no. SES-8012667, and is gratefully acknowledged. I thank Laurent Guy for research assistance, and Andrew Abel, Martin Baily, Fischer Black, Zvi Bodie, Benjamin Friedman, Daniel Holland, Robert McDonald, Julio Rotemberg, Richard Ruback, Paul Samuelson, Richard Schmalensee, Robert Shiller, Lawrence Summers, and two referees for helpful discussions and comments.

[1] Of course one could argue that no explanation is needed, i.e., the performance of the market was simply an "unlikely" realization of a stochastic process.

[2] It seems hard to believe that such a confusion would persist, particularly during a decade of high inflation. Also, as Lawrence Summers (1981a) points out, such confusion should also have led to declines in prices of owner-occupied housing.

[3] See Irwin Friend and Joel Hasbrouck (1982a) and Feldstein (1982). Patric Hendershott (1981) shows the effect is reduced considerably if debt and equity yields are made endogenous.

[4] Changes cited by Malkiel include the return of severe recessions (viewed as a thing of the past during the 1960's), a higher and more variable inflation rate (treated explicitly in this paper), and an increase in both the extent and unpredictability of government regulation.

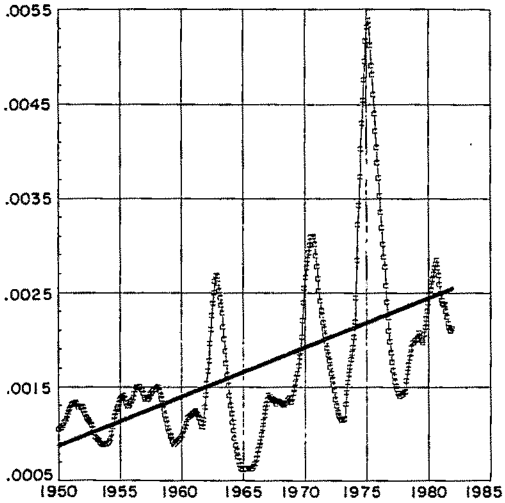


FIGURE 1. MONTHLY VARIANCE OF NOMINAL STOCK RETURNS (EXPONENTIALLY SMOOTHED)

Volatility in the firm's gross marginal return on capital comes from the stochastic nature of the instantaneous marginal product of capital (for example, crop harvests, worker productivity, and physical depreciation all have random components), and from the capital gains and losses caused by unforeseen events that alter the expected future flow of marginal revenue product from existing capital (for example, the effects of unanticipated regulatory change, exchange rate fluctuations that alter the competitive positions of goods produced abroad, etc.) Because the capital gains and losses are largely unrealized, the firm's gross marginal return on capital cannot be measured directly. However, its variance can be estimated indirectly from stock market data (assuming rational share valuation). As we will see, that variance has grown significantly, in a way consistent with Malkiel's suggestion that the business environment has become much more uncertain.

Increases in the expected rate of inflation have been accompanied by increases in the variance of that rate, and this can also affect the variance of stockholders' returns. First, inflation affects net real returns directly through the tax system, so that volatility of inflation causes volatility in these returns. Second, there is a well-known negative correlation between unanticipated inflation and stock returns.[5] I do not explain that correlation; as Fama (1981) has argued, it may in part be an indirect one occurring through correlations with real economic variables. However, it implies a negative correlation between unanticipated inflation and the gross marginal return on capital, so that volatility of inflation will be associated with volatility in that gross return.[6]

On the other hand, an increased volatility of inflation also increases the riskiness of nominal bonds. The relative size of the effect again depends on tax rates and other parameters, but we will see that overall it makes bonds relatively riskier, and should therefore increase share values.

Section I of this paper shows how investors' net real returns on equity and bonds depend on taxes, inflation, and the gross marginal return on capital. The specification of those returns extends Feldstein's (1980b) model so that risk is treated explicitly. In Section II, I discuss the data and parameter values, estimate the variance of the gross marginal return on capital and its covariance with inflation, and examine their behavior over time. In Section III, a simple partial equilibrium model is used to relate changes in the price of equity to changes in the mean and variance of inflation, and the mean and variance of the marginal return on capital. Section IV shows how changes in these means and variances over the past two decades can explain a good part of the behavior of share values.

[5] See Zvi Bodie (1976), Charles Nelson (1976), and Fama and G. William Schwert (1977).

[6] There are good reasons to expect this. For example, supply shocks create unanticipated inflation and at the same time reduce the current and expected future marginal products of capital. Also, as Richard Parks (1978) has shown, unanticipated inflation increases the dispersion of *relative* prices. This increases the dispersion of profits across firms (increasing risk for each firm), and if adjustment costs are significant, reduces expected profits overall. Related is Milton Friedman's (1977) suggestion that unanticipated inflation reduces economic efficiency by magnifying the distortions caused by regulation and long-term contracting, and reducing the signal-to-noise ratio in the messages transmitted by relative prices.

Before proceeding, the main argument of this paper can be illustrated with two simple regressions. Summers (1981a) used a "rolling ARIMA" forecast to generate an expected inflation series, $\pi^e(t)$, and then (using quarterly data for 1958–78) regressed the real excess returns (ER) on the NYSE Index on the change in expected inflation, $\Delta\pi^e(t)$. He obtained a negative coefficient for $\Delta\pi^e$, supporting his argument that increases in inflation cause decreases in share values.

I computed a similar series for $\pi^e(t)$, using annual averages of a rolling ARIMA forecast of monthly data.[7] The corresponding ordinary least squares (OLS) regression for the period 1958–81 is shown below (t-statistics are shown in parentheses):

$$\begin{aligned} ER = &.00335 - 3.615\Delta\pi^e \\ &(1.19) \quad (-2.29) \end{aligned}$$

$$R^2 = .193,\ SER = .0137,\ D\text{-}W = 1.93.$$

As in Summers' regression, the coefficient of $\Delta\pi^e$ is negative and significant. But now let us add another explanatory variable, the change in the variance of stock returns, $\Delta\sigma_s^2$:

$$\begin{aligned} ER = &.00258 + 1.481\Delta\pi^e - 8.936\Delta\sigma_s^2 \\ &(1.12) \qquad (0.76) \qquad (-3.48) \end{aligned}$$

$$R^2 = .488,\ SER = .0112,\ D\text{-}W = 1.85.$$

Observe that the coefficient of $\Delta\sigma_s^2$ is negative and highly significant, while the coefficient of $\Delta\pi^e$ is now insignificantly different from zero.[8] This simple regression suggests that increased risk and not increased inflation caused share values to decline. The analysis that follows explores that possibility.

I. Asset Returns

For simplicity, portfolio choice in this paper is limited to two assets, stocks and nominal bonds. I treat the rate of inflation as stochastic, so that the real returns on both of these assets are risky. Trading is assumed to take place continuously and with negligible transactions costs, and asset returns are described as continuous-time stochastic processes. As we will see, this provides a convenient framework for analyzing the effects of risk. In this section I derive and discuss expressions for investors' real after-tax asset returns. All of the parameters and symbols introduced here and throughout the paper are summarized in Appendix Table A1.

A. *The Return on Bonds*

I describe inflation and bond returns as in Stanley Fischer (1975). The price level follows a geometric random walk, so that the instantaneous rate of inflation is given by

$$(1) \qquad dP/P = \pi dt + \sigma_1 dz_1,$$

where $dz_1 = \varepsilon_1(t)(dt)^{1/2}$, with $\varepsilon_1(t)$ a serially uncorrelated and normally distributed random variable with zero mean and unit variance, that is, $z_1(t)$ is a Wiener process. Thus over an interval dt, expected inflation is πdt and its variance is $\sigma_1^2 dt$.

Bonds are short term, and yield a (guaranteed) gross nominal rate of return R. We can view this return as an increase in the nominal price of a bond, that is,

$$(2) \qquad dP_B/P_B = R dt.$$

The gross *real* return on the bond is therefore:[9]

$$(3) \quad d(P_B/P)/(P_B/P) = \left(R - \pi + \sigma_1^2\right) dt - \sigma_1 dz_1.$$

[7] At the end of each year, an ARIMA (4,0,0) model is estimated using monthly CPI data for the preceding 10 years, and is used to forecast the inflation rate for the next 12 months. I calculate π^e each year as an average over those 12 months. Note that ER and π^e are both measured as monthly rates.

[8] The results are qualitatively the same if my estimated series for the variance of the marginal return on capital is used instead of σ_s^2. The data are described in Section III. Fischer Black (1976) has shown earlier that stock returns tend to be contemporaneously negatively correlated with changes in price volatility.

[9] Equation (3) is obtained by use of Ito's Lemma. Fischer derives equation (3), and also provides a brief introduction to Ito processes such as (1), and Ito's Lemma and its use. Observe that the greater the variance σ_1^2 of the inflation rate, the greater the expected real return on the bond. This is just a consequence of Jensen's inequality; the bond's real price P_B/P is a convex function of P.

Interest payments are taxed as income, so the investor's net nominal return on bonds is $(1-\theta)R\,dt$, where θ is the personal income tax rate. The *net real return* on bonds over an interval dt is therefore:

$$(4)\quad \xi_b = \left[(1-\theta)R - \pi + \sigma_1^2\right] dt - \sigma_1\, dz_1$$
$$= r_b\, dt - \sigma_1\, dz_1.$$

This characterization of bond returns contains the simplifying assumption that stochastic changes in the price level are serially uncorrelated.[10] If inflation actually followed (1), the real return on long-term bonds would be no riskier than that on short-term bonds.[11] In reality, stochastic changes in the price level are autocorrelated, so that long-term bonds are indeed riskier. My model could be expanded by adding long-term bonds as a third asset and allowing for autocorrelation in price changes, but the added complication would buy little in the way of additional insight, and would not qualitatively change any of the basic results.

B. *The Return on Stocks*

To derive an expression for investors' net real return on stocks, I begin with a description of the firm's gross marginal return on capital. Following Feldstein (1980b), I then introduce the effects of inflation and taxes.

Over a short interval of time, dt, the gross real return to the firm from holding a marginal unit of capital will consist of two components: the instantaneous marginal product of the unit, and the instantaneous change in the present value of the expected future flow of marginal product. This second component is just a capital gain or loss. However, it will generally be an *unrealized* capital gain or loss, so that the firm's gross marginal return on capital is not an accounting return.

Both components will be in part stochastic. The current marginal product of capital will have a stochastic element arising from random shocks in the production process: the weather in farming, random discovery rates in response to natural resource exploration, strikes, random week-to-week fluctuations in labor productivity, etc. Capital gains and losses are almost entirely stochastic, and occur when unforeseen events alter the expected value of the *future* flow of marginal product: for example, an OPEC oil shock that reduces the value of factories producing large cars while raising the value of drilling equipment, an exchange rate fluctuation that gives certain domestically produced goods a competitive advantage or disadvantage, a regulatory change that makes some existing capital obsolete or raises the cost of using it, and so forth.

On an aggregate basis, it is reasonable to assume that the stochastic part of the real gross marginal return on capital is normally distributed. We can then write that return as

$$(5)\qquad m = \alpha\, dt + \sigma_2\, dz_2,$$

where α is the expected return (largely the expected current marginal product). As Fama (1981) and others have stressed, this return is likely to be negatively correlated with the rate of inflation.[12] The magnitude and significance of this correlation will be addressed shortly; here I simply denote $E(dz_1\, dz_2) = \rho\, dt$.

Although both the current and expected future marginal products of capital contribute to the stochastic term in equation (5), most of the variance is due to the capital gain component.[13] Since these capital gains

[10] I also assume the nominal interest rate is nonstochastic, so all of the risk from holding bonds (apart from default risk) comes from uncertainty over inflation. As Fama (1975) has shown, this assumption is roughly consistent with the historical data.

[11] Note that a pure discount bond that pays $1 at time T has a present value at time t of

$$\exp\left[-\int_t^T \left(R - \pi + \sigma_1^2\right) dt + \int_t^T \sigma_1\, dz_1\right],$$

and therefore a real instantaneous return of $(R - \pi + \sigma_1^2)\,dt - \sigma_1\, dz_1$, as in equation (3) for the short-term bond.

[12] Note that this is apart from the effects of inflation on investors' net real return that are brought about by the tax system, as explained by Feldstein (1980a, b).

[13] It is shown in Section II that the variance of the current marginal product of capital only accounts for about 2 percent of the variance of m.

and losses are largely unrealized, they are not taxed directly. However they are taxed indirectly in that the corresponding *future* marginal products are taxed. We can therefore treat the corporate income tax as applying to both the deterministic and stochastic components of m.[14] Letting τ_s be the statutory corporate income tax rate, and τ_e the *effective* corporate income tax rate ($\tau_e < \tau_s$ because of accelerated depreciation and the investment tax credit), and denoting corporate borrowing per unit of capital by b, the firm's *net* real return on capital in the *absence of inflation* is then $(1-\tau_e)\alpha\,dt - (1-\tau_s)bR\,dt + (1-\tau_e)\sigma_2\,dz_2$.

Following Feldstein (1980b), we can adjust this net return for the effects of inflation. First, inflation reduces the real value of the firm's debt, so that the net after-tax cost of borrowing is $(1-\tau_s)bR\,dt - b(dP/P)$.[15] Second, because the value of depreciation allowances is based on original or "historic" cost, inflation reduces the real value of depreciation and increases real taxable profits. I use Feldstein's linear approximation that a 1 percent increase in the price level reduces net profits per unit of capital by an amount λ. Then letting q denote the price of a share (representing a unit of capital), the firm's real net earnings per dollar of equity over an interval dt, ψ_s, is given by

$$(6)\quad (1-b)q\psi_s = (1-\tau_e)\alpha\,dt - (1-\tau_s)bR\,dt + (1-\tau_e)\sigma_2\,dz_2 + (b-\lambda)(dP/P).$$

Substituting equation (1) for dP/P, we then have

$$(7)\quad (1-b)q\psi_s = \left[(1-\tau_e)\alpha - (1-\tau_s)bR + (b-\lambda)\pi\right]dt + (b-\lambda)\sigma_1\,dz_1 + (1-\tau_e)\sigma_2\,dz_2.$$

Now consider the after-tax return to investors. Let d be the fraction of net earnings paid out as dividends, and θ_c the *effective* tax rate on capital gains. Then in the absence of inflation, investors' net real return per dollar of equity would be $\psi_s[(1-\theta)d + (1-\theta_c)(1-d)q]$. Inflation creates nominal capital gains at a rate $(dP/P)q$ per share, or $(dP/P)q(1-b)$ per unit of capital. Thus investors' net real return per dollar of equity, ξ_s is given by

$$(8)\quad \xi_s = \psi_s\left[(1-\theta)d + (1-\theta_c)(1-d)q\right] - \theta_c(dP/P).$$

Again, we can substitute for dP/P, and for ψ_s. Letting

$$(9)\quad a = \left[(1-\theta)d + (1-\theta_c)(1-d)q\right]/(1-b)q,$$

the net return is

$$(10)\quad \xi_s = \left\{a\left[(1-\tau_e)\alpha - (1-\tau_s)bR + (b-\lambda)\pi\right] - \theta_c\pi\right\}dt + \left[a(b-\lambda) - \theta_c\right]\sigma_1\,dz_1 + a(1-\tau_e)\sigma_2\,dz_2 = r_s\,dt + s_1\,dz_1 + s_2\,dz_2.$$

C. *Inflation and Asset Returns*

Feldstein (1980a, b) has argued that increased inflation has reduced the expected real net return to investors from holding stocks, thereby depressing share values. As Irwin Friend and Joel Hasbrouck (1982a) have shown, that argument depends on tax rates and other parameter values. It also depends on the way the nominal interest rate R changes in response to changes in the expected inflation rate π. The conventional wisdom is $dR/d\pi = 1$, at least in long-run equilibrium. As Summers (1983) shows, *in theory $dR/d\pi$* should be about 1.3 (if savings are interest inelastic) because of the taxation of *nominal* interest payments. Summers provides convincing evidence that $dR/d\pi$ has historically been much less than 1 (at most

[14] This is an approximation, first because losses may more than offset taxable profits, and second because depreciation allowances are calculated on an *ex ante* basis. For an analysis of the risk-shifting effects of the corporate income tax, see Jeremy Bulow and Summers (1982).

[15] I ignore non-interest-bearing monetary assets, which are small relative to interest-bearing debt.

about 0.6) even in the long run. I take the true value of $dR/d\pi$ to be an unresolved empirical question,[16] and denote it by the *parameter* R_π. We can then examine how changes in π should affect share values for alternative values of R_π.

To see how inflation affects asset returns, numerical values are needed for the tax rates θ, θ_c, and τ_s, as well as the parameters b, λ, and d. All of the parameters are discussed in Appendix A, and reasonable values are $\theta = .30$, $\theta_c = .05$, $\tau_s = .48$, $b = .30$, $\lambda = .26$, and $d = .43$. Setting $q = 1$, we have $a = 1.2$.

As can be seen from equations (4) and (10), an increase in π reduces investors' expected return on equity as long as R_π is positive, but it also reduces the expected return on bonds as long as $R_\pi < 1.43$. What is relevant is the differential effect on expected stock returns vs. bond returns. The parameter values above imply that

$$(11)\quad (d/d\pi)[E(\xi_s)-E(\xi_b)]/dt = -.887R_\pi + .998,$$

which is positive as long as $R_\pi < 1.13$. Since most estimates put R_π below 1.13, it seems doubtful that increases in expected inflation depresssed share values by differentially reducing the expected return on equity; in fact they could have worked to *increase* share values.

Asset demands also depend on the variances of these net returns. As shown shortly, σ_2^2, the variance of the gross marginal return on capital m, has increased significantly over time, and this *can* explain much of the decline in share values. The variance of inflation, σ_1^2, also increased in the 1970's. This increased the variance of bond returns, but to what extent did it contribute to the variance of investors' net return on equity? Observe from equation (10) that the variance of that return is

$$(12)\quad (1/dt)\mathrm{Var}(\xi_s) = [a(b-\lambda)-\theta_c]^2\sigma_1^2 + a^2(1-\tau_e)^2\sigma_2^2 + 2a[a(b-\lambda)-\theta_c](1-\tau_e)\sigma_1\sigma_2\rho.$$

Using the parameter values from above, this becomes

$$(13)\quad (1/dt)\mathrm{Var}(\xi_s) = .000004\sigma_1^2 + .518\sigma_2^2 - .0029\rho\sigma_1\sigma_2.$$

Thus any increases in the variance of inflation would have had a negligible direct effect on the variance of the net real return on equity.[17] Of course increases in σ_1^2 could have also affected share values by shifting the demand for bonds, but as shown in Section III, the magnitude of any such effect is small.

II. Inflation and the Marginal Return on Capital over Time

The variance of the real gross marginal return on capital, σ_2^2, and its coefficient of correlation with inflation, ρ, cannot be observed directly. However they can be inferred from σ_s^2, the variance of nominal stock returns, Ω_{sp}, the covariance of nominal stock returns with inflation, the inflation variance σ_1^2, and the tax and financial parameters. Here I estimate these variances and covariance, infer values of σ_2^2 and ρ, and examine their behavior over time.

I use a crude method to estimate the mean and variance of the monthly inflation rate, and the monthly variance σ_s^2 and covariance Ω_{sp}. Assuming the true values of these parameters are slowly varying over time (i.e., are roughly constant over intervals up to a year, but may vary over periods of several years), I compute a moving 13-month centered sample mean, and sample variances and covariance. This yields estimates that are rough, but at least as accurate as available estimates of the various tax and financial parameters.

The mean inflation rate π is computed as a moving, 13-month centered sample mean,

[16] In addition to Summers (1983), see Fama (1975), and Nelson and Schwert.

[17] This is because $a(b-\lambda) \approx \theta_c$. Of course the variance of inflation could have had a significant *indirect* effect on the variance of stock returns by partially "explaining" increases in σ_2^2, the variance of m (see fn. 6). I account for inflation when estimating σ_2^2 in Section II.

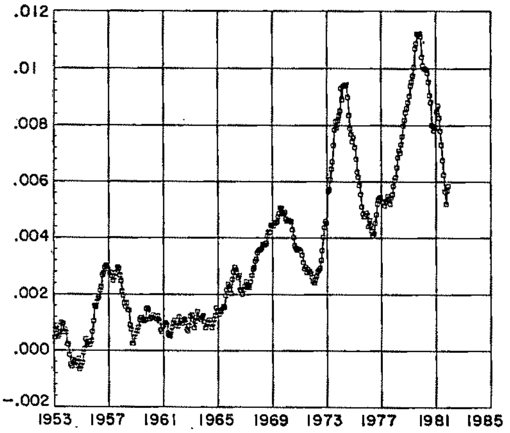


FIGURE 2. MEAN INFLATION RATE (MONTHLY)

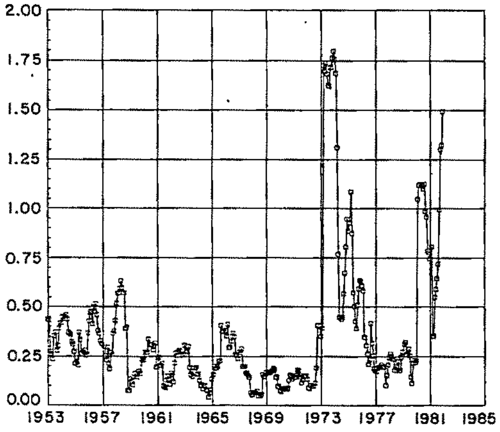


FIGURE 3. MONTHLY VARIANCE OF INFLATION

using the *CPI* as the price index:

$$(14) \qquad \pi_t = \sum_{j=-6}^{6} \Delta \log P_{t+j}/13.$$

Similarly, a monthly sample variance is computed for σ_1^2:

$$(15) \qquad \sigma_1^2(t) = \sum_{j=-6}^{6} (\Delta \log P_{t+j} - \pi_t)^2/12.$$

Trends in π_t and $\sigma_1^2(t)$ over the period 1953–81 are illustrated in Figures 2 and 3.[18] Observe the clear upward trend in π from 1965 to 1981; the average annual inflation rate for 1953–68 was 2 percent, compared to 9 percent for 1973–81. Increases in the variance of inflation were roughly confined to the oil shocks and recessions of 1973–75 and 1979–82; σ_1^2 had an average 1953–68 value of 2.6×10^{-6}, and an average 1973–81 value of 6.9×10^{-6}.

Monthly total (nominal) returns data for the New York Stock Exchange Index were obtained from the CRISP tape. The sample variance σ_s^2 was computed using a constant value of 0.71 percent for the monthly expected return:[19]

$$(16) \qquad \sigma_s^2(t) = \sum_{j=-6}^{6} (x_{t+j} - .0071)^2/12,$$

where x_t is the logarithmic return at time t. (See Figure 1.) Finally, the covariance Ω_{sp} is computed as

$$(17) \qquad \Omega_{sp}(t) = \sum_{j=-6}^{6} (x_{t+j} - .0071) \times (\Delta \log P_{t+j} - \pi_t)/12.$$

To infer values of σ_2^2 and ρ from σ_s^2, Ω_{sp}, σ_1^2, and the various parameters, note that *nominal* net earnings per dollar of equity, ψ_s^n, are given by

$$(18) \qquad \psi_s^n = \psi_s + dP/P + \psi_s(dP/P).$$

[18] Inflation was extremely volatile during the Korean War, exceeding 8 percent during the first 6 months after the outbreak of the war in July 1950, and dropping to less than 1 percent in 1952.

[19] This is the mean return for the period 1960–81, and is close to Robert Merton's (1980) estimate of 0.87 percent obtained using data for 1926–78. In fact, the expected return on the market will change if corporate tax rates change, if π changes, or if the expected real gross marginal return on capital changes. However, as Merton (1980) shows, even if we took this expected return to be *zero*, it would bias our estimates of σ_s^2 only slightly. Similarly, using a moving sample mean (as in the computation of σ_1^2) makes a negligible difference; for 1948–81, the monthly series for σ_s^2 computed in this way has a correlation coefficient of .981 with the corresponding series computed from equation (16). Finally, note that σ_s^2 should ideally be computed using *daily* data, but such data are not available before 1962.

Substituting equation (1) for dP/P and equation (7) for ψ_s yields

$$(19)\quad (1-b)q\psi_s^n = \big[(1-\tau_e)\alpha-(1-\tau_s)bR + (b-\lambda)\pi+(1-b)q\pi+(b-\lambda)\sigma_1^2 + (1-\tau_e)\rho\sigma_1\sigma_2\big]\,dt + \big[(b-\lambda)+(1-b)q\big]\sigma_1\,dz_1 + (1-\tau_e)\sigma_2\,dz_2.$$

It will be assumed that shares are rationally valued,[20] so that $\sigma_s^2 = (1/dt)\mathrm{Var}(\psi_s^n)$, and $\Omega_{sp} = (1/dt)\mathrm{Cov}(\psi_s^n, dP/P)$. Thus,

$$(20)\quad \sigma_s^2 = \Big\{(1-\tau_e)^2\sigma_2^2 + \big[(b-\lambda) + (1-b)q\big]^2\sigma_1^2 + 2(1-\tau_e)\big[(b-\lambda) + (1-b)q\big]\rho\sigma_1\sigma_2\Big\}\big/(1-b)^2q^2,$$

$$(21)\quad \Omega_{sp} = \Big\{\big[(b-\lambda)+(1-b)q\big]\sigma_1^2 + (1-\tau_e)\rho\sigma_1\sigma_2\Big\}\big/(1-b)q.$$

These equations are solved simultaneously for σ_2^2 and ρ, given values for the tax and financial parameters. Most of these latter parameters have been estimated by others, or can be roughly calculated in a straightforward way. Values for all of them are discussed in Appendix A, and are summarized in Appendix Table A1.

The calculated series for σ_2^2 is shown in Figure 4, together with a fitted trend line.[21] Observe that movements of σ_2^2 closely parallel those of σ_s^2, with a clear positive trend beginning about 1960. The average value of σ_2^2 for 1953–68 was .0017, and for 1973–81, .0036, more than doubling. σ_2^2 was sharply higher during the oil and agricultural price of shock of 1974 and recession of 1975, but a strong positive trend remains even if these years are excluded; its average value for 1976–81 was .0027, a large increase from the 1950's and 1960's.

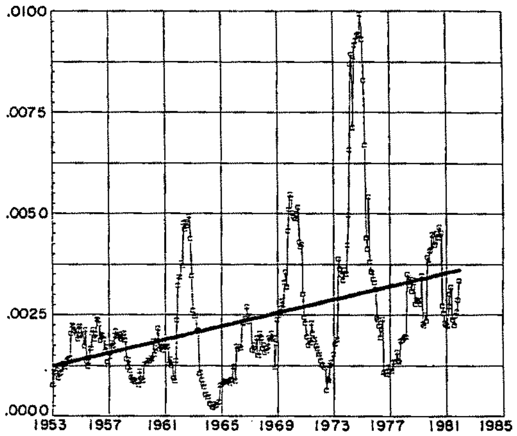


FIGURE 4. MONTHLY VARIANCE OF MARGINAL RETURN ON CAPITAL

It is also interesting to compare σ_2^2 with the variance of the marginal product of capital. Daniel Holland and Stewart Myers (1980) estimated the latter to be about .000576 on an annual basis, or 4.8×10^{-5} on a monthly basis. This is roughly 2 percent of my average estimates of σ_2^2, confirming that most of the variance of the marginal return on capital is due to capital gains and losses.

The monthly estimates of the correlation coefficient ρ fluctuate considerably over time, and are shown exponentially smoothed ($\tilde{\rho}_t = .1\rho_t + .9\tilde{\rho}_{t-1}$) in Figure 5. Observe that ρ was not always negative. One might expect ρ to increase or turn positive during periods when economic fluctuations are driven by demand shocks. This is consistent with its behavior during the 1950's, and with a 1953–68 average value of $-.092$, as compared to a 1973–81 average value of $-.244$.

To summarize, over the past fifteen or twenty years there have been large increases in π, σ_1^2, and σ_2^2, and a decrease in ρ. I should add to this that there is evidence from

[20] As Robert Shiller (1981) has shown, the validity of this assumption is questionable.

[21] The trend line is (t-statistics in parentheses):

$$\bar{\sigma}_2^2(t) = \underset{(3.31)}{7.51\times10^{-4}} + \underset{(8.56)}{6.88\times10^{-6}}t, \;(R^2 = .157).$$

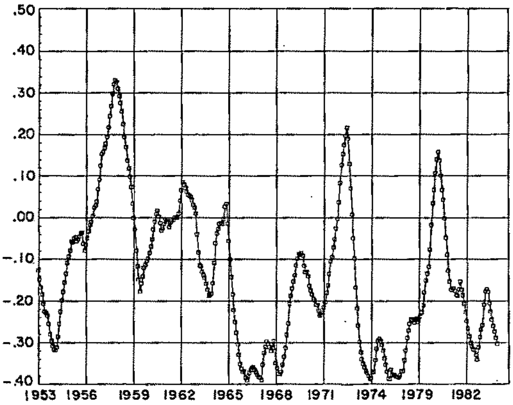


FIGURE 5. CORRELATION COEFFICIENT ρ (SMOOTHED)

the research of others that α, the expected gross marginal return on capital, has declined from about .12 to about .10 at an annual rate. In order to determine the implications of these trends for share values, we need a model of asset demands.

III. Asset Demands and Share Values

Given equations (4) and (10) for the asset returns ξ_b and ξ_s, we can use the solution to the investor's consumption-portfolio problem to determine asset demands, and thereby determine how share values change in response to changes in π, σ_1^2, σ_2^2, etc. To do this, assume that future income streams are certain and can be borrowed against, and can therefore be capitalized in initial wealth. Also assume that investors have constant relative risk-aversion utility of consumption C.[22] The consumption-portfolio problem is then:

$$(22)\quad \max_{C,\beta} E_0 \int_0^{\infty} \frac{1}{1-\gamma}\left(C^{1-\gamma}+\gamma-2\right)e^{-\delta t}\,dt,$$

subject to

$$(23)\quad dW=\left[\beta(r_s-r_b)W+r_bW-C\right]dt + \beta\left[s_2\,dz_2+(\sigma_1+s_1)\,dz_1\right]W - \sigma_1 W\,dz_1,$$

where W is wealth, β is the fraction of wealth invested in stocks, $\gamma > 0$ is the index of relative risk aversion, and r_s, r_b, s_1 and s_2 are defined in equations (4) and (10).

This is similar to the consumption-portfolio problem in Merton (1971), except that both assets are risky. The solution is (see Appendix B)

$$(24)\quad \beta^* = (r_s-r_b)/\gamma\Sigma_{12}^2 + \sigma_1\Sigma/\Sigma_{12}^2,$$

where

$$(25)\quad \Sigma_{12}^2 \equiv s_2^2 + 2\rho s_2(\sigma_1+s_1) + (\sigma_1+s_1)^2 = (1/dt)\mathrm{Var}(\xi_s-\xi_b),$$

and

$$(26)\quad \sigma_1\Sigma \equiv \sigma_1(\sigma_1+s_1+\rho s_2) = (1/dt)\mathrm{Var}(\xi_b)-(1/dt)\mathrm{Cov}(\xi_s,\xi_b).$$

Observe that we can also write the portfolio rule (24) in terms of the deviation from equal shares ($\beta = 1/2$):

$$(27)\quad \beta^* = 1/2 + \frac{E(\xi_s-\xi_b)-\gamma\left[\mathrm{Var}(\xi_s)-\mathrm{Var}(\xi_b)\right]}{\gamma\,\mathrm{Var}(\xi_s-\xi_b)}.$$

Thus the share of wealth held in stocks depends in an intuitively appealing way on the relative expectations and variances of the returns. The amount by which that share exceeds one-half is proportional to the difference in the expected returns less the difference in the variances of the returns (adjusted by the index of risk aversion).[23] Finally, note that this portfolio rule is also the one that maximizes the weighted sum $E(\xi_p)-\gamma\mathrm{Var}(\xi_p)/2$, where $\xi_p=\beta\xi_s+(1-\beta)\xi_b$ is the portfolio return.

To determine the effects of changes in π, σ_1^2, σ_2^2, etc. on share values, I use a partial

[22] Friend and Marshall Blume (1975) provide empirical support for this assumption.

[23] Because the *real* gross marginal product of capital is correlated with inflation, there is no feasible value of ρ that makes stocks and bonds perfect substitutes in this model. The assets are perfect substitutes if $\Sigma_{12}^2 = 0$, but this would require $\rho = [s_2^2+(s_1+\sigma_1)^2]/2s_2(s_1+\sigma_1) < -1$.

equilibrium framework and assume a fixed quantity of stock s, and fixed aggregate wealth. Let ϕ denote some parameter of interest. Remember that $\beta^* = \beta^*(q, \phi)$, where q is the share price, so that

$$(28)\quad \frac{dq}{d\phi} = \frac{W}{s}\frac{d\beta^*}{d\phi} = \frac{q}{\beta^*}\left(\frac{\partial\beta^*}{\partial\phi} + \frac{\partial\beta^*}{\partial q}\frac{dq}{d\phi}\right)$$

$$\text{or}\quad d\log q/d\phi = \frac{\partial\beta^*/\partial\phi}{\beta^* - q(\partial\beta^*/\partial q)}.$$

Formulas for $d\log q/d\pi$, $d\log q/d\sigma_1^2$, $d\log q/d\alpha$, $d\log q/d\sigma_2^2$, and $d\log q/d\gamma$ are given in Appendix C. The numerical values of these derivatives will of course depend on the values of π, σ_1^2, σ_2^2, etc., as well as the tax and financial parameters. Here I calculate values for the derivatives using the following average values for the 1965–81 time period: $\bar{\pi} = .00539$, $\bar{R} = .00625$, $\bar{\sigma}_1^2 = 4.6 \times 10^{-6}$, $\bar{\sigma}_2^2 = .0030$, $\bar{\rho} = -.22$, $\bar{\alpha} = .00858$. (Note that these are *monthly* means and variances.) The values of the tax and other parameters are those listed in the summary Appendix Table A1, and discussed in Appendix A.[24]

The derivatives (28) are calculated around a base value of β^*. We take β^* to be the ratio of the value of equity to the value of equity plus debt, both long- and short-term. (To keep the model simple, I am assuming separability of such assets as human capital, housing, land, money, etc., and ignoring the risk differentials across various debt instruments.) The derivatives also depend on the value of γ, the index of relative risk aversion. Taking the other parameter values as given, one can choose γ so that the calculated value of β^* is equal to 0.67, its average for the 1965–81 period as given by the National Balance Sheets. That value of γ is 5.8, which may appear large, but is consistent with Friend and Blume's estimates showing γ to be "in excess of two," Friend and Hasbrouck's (1982b) estimate showing γ to be about 6, and Grossman and Shiller's finding that γ appears to be about 4. I take the true value of γ to be an unresolved empirical question, and calculate numerical values of the derivatives for alternative values of γ. These are shown in Table 1.

Observe that the sign of $d\log q/d\pi$ depends on the value of R_π. Changes in π affect only expected returns and not their variances or covariance, so from equation (11), $d\log q/d\pi > (<)\ 0$ if $R_\pi < (>)\ 1.13$. The numbers in Table 1 indicate that changes in π should have a small effect on share values—unless R_π is around .6 or less, as Summers' (1983) findings indicate may be the case. If $R_\pi = 1.0$, an increase in the expected annual inflation rate from 5 to 10 percent ($\Delta\pi = .0039$) implies a 7 percent *increase* in q if $\gamma = 5$, and a 10 percent increase if $\gamma = 3$. However, if $R_\pi = .6$, the result is a 30 percent increase in q if $\gamma = 5$, and a 44 percent increase if $\gamma = 3$.

The sign of $d\log q/d\sigma_1^2$ depends on γ (see equation (A8) in the Appendix). For our parameter values, $d\log q/d\sigma_1^2 > (<)\ 0$ if $\gamma > (<)\ .987$. An increase in σ_1^2 increases the relative riskiness of bonds, but it also increases their expected return (see equation (4)). If γ is small enough, the second effect increases the demand for bonds more than the first reduces it, so that q falls. Since γ is probably greater than one and possibly greater than four, we would expect $d\log q/d\sigma_1^2 > 0$. However, this derivative is small given the average size of σ_1^2. In 1973–75, σ_1^2 roughly quadrupled from a pre-OPEC average value of about 2×10^{-6}. (See Figure 3.) This ($\Delta\sigma_1^2 = 6 \times 10^{-6}$) implies only a 0.4 percent increase in q if $\gamma = 5$ and a 0.3 percent increase if $\gamma = 3$.

As I have noted, there is some evidence that α, the expected real gross marginal return on capital, declined somewhat during the 1970's, perhaps from .12 to .10 at an annual rate.[25] This decline in α ($\Delta\alpha = .0015$)

[24] These values of ρ, σ_1^2, σ_2^2, R, π, α, and the tax and financial parameters imply an expected after-tax return to the firm $E(\psi_s)$, of .629 percent monthly, or 7.8 percent annually. This is well within the range of estimates of the after-tax real rate of return.

[25] The evidence is mixed. See Feldstein and Summers (1977), and Feldstein, James Poterba, and Louis Dicks-Mireaux (1981). Holland and Myers show a larger decline, from about 15 to 11 percent, but do not adjust for cyclical variation.

TABLE 1—VALUES OF DERIVATIVES AT POINT OF MEANS

	$d \log q/d\pi$						
γ	$R_\pi = 1.3$	$R_\pi = 1.0$	$R_\pi = 0.6$	$d \log q/d\sigma_1^2$	$d \log q/d\alpha$	$d \log q/d\sigma_2^2$	$d \log q/d\gamma$
.2	−115.0	82	345	−591	533	−51	−22.3
.5	−94.3	67.2	283	−299	436	−105	−7.29
1	−72.4	51.6	217	6	335	−161	−2.80
2	−49.4	35.3	148	326	229	−220	−0.96
3	−37.5	26.8	113	493	174	−251	−0.48
4	−30.3	21.6	90.7	594	140	−270	−0.29
5	−25.3	18.1	75.9	663	117	−282	−0.20
6	−21.8	15.5	65.3	712	101	−292	−0.14
7	−19.1	13.6	57.3	750	89	−298	−0.11
8	−17.0	12.1	51.1	779	79	−303	−0.08
10	−14.0	10.0	41.9	822	65	−312	−0.05
∞	0	0	0	1017	0	−348	0

would imply an 18 percent decline in q if $\gamma = 5$, and a 26 percent decline if $\gamma = 3$. This is clearly a large effect, so even a *perceived* (as opposed to actual) decline in α could explain a significant amount of the market's performance.

As shown in the preceding section, there is evidence that over the past two decades σ_2^2 has more than doubled from an average 1953–68 value of about .0017 (monthly) to an average 1973–81 value of .0036. This increase in σ_2^2 would imply a *54 percent decline in q* if $\gamma = 5$, a 48 percent decline if $\gamma = 3$, and a 31 percent decline if $\gamma = 1$. If σ_2^2 (or investors' estimates of σ_2^2) indeed doubled, this would explain a large part of the market's decline for any reasonable value of γ.

Finally, observe that an increase in γ would also have a large negative effect on share values. For example, an increase in γ from 4 to 5 would imply a 29 percent decline in q. However, I have no evidence that γ has changed over time one way or the other, so the focus in the next section is only on changes in π, σ_1^2, α, and σ_2^2.

IV. Explaining the Decline in Share Values

In Section II, I observed differing patterns of change in π, α, σ_1^2, and σ_2^2, but in all cases there have been reasonably clear shifts from the period 1953–68 to the period 1973–81. To summarize those shifts, π increased from a 1953–68 average value of 2 percent annually to a 1973–81 average value of 9 percent annually ($\Delta\pi = .0056$ on a monthly basis), α fell from about a 12 percent annual rate to about a 10 percent annual rate ($\Delta\alpha = -.0015$), and the average monthly variances σ_1^2 and σ_2^2 increased from 2.6×10^{-6} to 6.9×10^{-6} and .0017 to .0036, respectively. Table 2 and Figure 6 show the effects of these changes on share values, individually and in combination, as a function of γ, and assuming $R_\pi = 1.0$.

Observe that the relative importance of changes in π, α, and σ_2^2 depends on the index of risk aversion γ. For very small values of γ, changes in π and α have large effects on q, but if γ exceeds 2 or 3, the change in σ_2^2 clearly dominates. The model developed in this paper suggests a value of γ around 5 or 6, and the work of Friend and Blume, Friend and Hasbrouck (1982b), and Grossman and Shiller support values that are at least in excess of 2.

The increases that occurred in σ_1^2 should have had a negligible direct effect on share values. Of course, I have ignored the fact that unanticipated price changes are autocorrelated, which in the 1970's caused a pronounced increase in the riskiness of longer-term bonds. For example, for 1973–81 the variance of the real return on five-year government bonds was about 50 times as large as that on one-month Treasury bills.[26]

[26] See Bodie, Alex Kane, and Robert McDonald (1983).

TABLE 2—CHANGES IN SHARE VALUES (Percentage change, assumes $R_\pi = 1.0$)

γ	$\Delta\sigma_2^2 = .0019$	$\Delta\sigma_1^2 = 4.3\times10^{-6}$	$\Delta\pi = .0056$	$\Delta\alpha = -.0015$	Δq_{TOTAL}
.2	−9.7	−0.3	45.9	−80.0	−44.1
.5	−20.0	−0.1	37.6	−65.4	−47.9
.8	−26.9	0.0	32.6	−56.3	−50.6
1	−30.6	0.0	28.9	−50.2	−51.9
2	−41.8	0.1	19.8	−34.4	−56.3
3	−47.7	0.2	15.0	−26.1	−58.6
4	−51.3	0.3	12.1	−21.0	−59.9
5	−53.6	0.3	10.1	−17.6	−60.8
6	−55.5	0.3	8.7	−15.2	−61.7
7	−56.6	0.3	7.6	−13.3	−62.0
8	−57.6	0.3	6.8	−11.9	−62.4
10	−59.3	0.4	5.6	−9.8	−63.1
∞	−66.1	0.4	0	0	−69.9

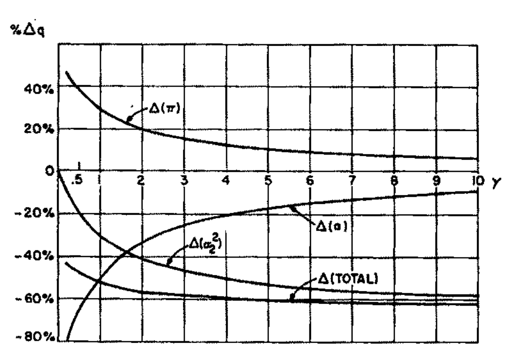


FIGURE 6. CONTRIBUTIONS OF CHANGES IN π, α, AND σ_2^2 TO CHANGE IN SHARE VALUES

We can roughly (and conservatively) account for this in my model by scaling up σ_1^2 and $\Delta\sigma_1^2$ by a factor of 50. The value of $d\log q/d\sigma_1^2$ in Table 1 then falls by about a factor of 5 if $\gamma = 3$, and a factor of 3 if $\gamma = 5$ (the other derivatives remain virtually unchanged), so that the implied increase in q would still not exceed 5 percent. Thus even if most debt were longer term, accounting for autocorrelation would not change this result.

If one accepts that $\gamma > 2$, then my results point to increased capital risk as the major cause of the decline in share values. If one believes instead that $\gamma < 2$, then the decline in the expected return α stands out as the major explanatory factor, although increased capital risk is still important. As for the increase in expected inflation, this by itself should have *increased* share values, unless R_π is larger than recent estimates indicate. Of course, one might argue that increased inflation has been a cause of both the decline in the expected return and the increase in risk, in which case it would have indirectly contributed to lower share values.

V. Concluding Remarks

I have found that much of the decline in share values can be attributed to the behavior of the gross marginal return on capital. Others have shown that the expectation of that return has fallen (although there is some controversy over how much), and this paper has shown that its variance has approximately doubled. The relative importance of these two changes depends on investors' index of risk aversion; if that index exceeds 2, increased risk is the dominating factor.

These results are based on a very simple model of asset returns, asset demands, and share price determination. Some of the model's limitations have already been mentioned: the reliance on rational share valuation for the estimation of σ_2^2, the assumption of a deterministic income stream in the consumption-portfolio problem, and the inclusion of only two assets in investors' portfolios. Even more limiting is the use of a partial equilibrium framework, which ignores the fact that as q falls the capital stock will begin to fall, and the expected return α will begin to rise, pushing q back up. Accounting for this would reduce the magnitude of the derivatives in Table 1, at least for the longer term.

Another issue that requires more attention is how investors' perception of capital risk should be measured. We estimate capital risk from the sample variance of stock market returns, but the use of survey data might yield better estimates of investors' perceptions of that risk. For example, if investors believe that there is a nonnegligible probability of economic catastrophe, capital would be quite risky, even if stock returns are not very volatile.

Finally, I have implicitly assumed that increases in the estimated values of σ_2^2 reflect actual increases in the riskiness of capital as an aggregate. Other interpretations are possible once we recognize that the capital stock is heterogenous. For example, suppose technological change caused the expected return on *riskier capital* (say the shares of high-tech growth firms) to rise. The demand for these riskier stocks would then rise, increasing the variance of the returns on a value-weighted aggregate stock index, and possibly causing share values on average to rise.

Appendix

A. *Tax and Financial Parameters*

The relevant tax parameters are the statutory and effective corporate tax rates τ_s and τ_e (the latter accounts for accelerated depreciation and the investment tax credit), the marginal personal tax rate θ, and the effective tax rate on capital gains θ_c. Over 1960–82, the statutory corporate tax rate has varied from a high of .528 in 1968–69 to a low of .46 since 1979. I use a constant average value of .48. Because the effects of inflation are explicitly included in my model, to avoid double counting we must remove those effects when computing the effective tax rate τ_e. To do this I utilize Feldstein and Summers' (1979) estimates of the excess tax due to inflation. As can be seen in Table A2, the effective rate τ_e has fluctuated, but has had an average value of .41 for 1962–77. Given the lower values for 1975–77, I use a constant value of .40 in the model. Feldstein and Summers (1979) also estimate the effective tax rate on capital gains and find it to be about .05, the value that I use. Finally, I use a value of .30 for the personal tax rate θ.

A rough value for b, corporate borrowing per unit of capital, can be computed by dividing the total debt of nonfinancial corporations by the total value of their capital, using data from the *National Balance Sheets*. That ratio was .26 to .28 during 1962–66, but from 1967 to 1981 it remained between .29 and .31. I use a constant average value of .30 for b. Similarly, a rough value for d, the fraction of net earnings paid out as dividends, can be computed by dividing aggregate dividends by aggregate after-tax profits, using data from the *Survey of Current Business*. That ratio has declined from about .48 in 1962 to about .34 in 1980; I use a constant average value of .43. Finally, a value is needed for λ, which measures the reduction in net profits per unit of capital resulting from a 1 percent increase in the price level. I use a value of 0.26, as estimated by Feldstein (1980b).

One parameter remains, and it is an important one: α, the expected real gross marginal return on capital. There has been some debate over whether α has been falling during the past decade or two, and the evidence is mixed, particularly if α is measured on a cyclically adjusted basis. Using the cyclically adjusted estimates of α made by Feldstein, Poterba, and Dicks-Mireaux, the average value for 1962–79 is .11 annual (.00858 monthly). Those estimates suggest that α has declined from .12 to .10, and I calculate the effect of such a decline on share values.

B. *Optimal Consumption-Portfolio Rule*

To solve the consumption/portfolio problems of equations (22) and (23), write the value function $V(W)$:

$$\text{(A1)} \quad V = \max_{C,\beta} E_t \int_t^\infty \frac{1}{1-\gamma}\left(C^{1-\gamma} + \gamma - 2\right) e^{-\delta(\tau - t)}\, d\tau,$$

TABLE A1—SUMMARY OF PARAMETERS AND SYMBOLS

Parameter	Definition	1965–81 Mean Value[a]
a	$[(1-\theta)d+(1-\theta_c)(1-d)q]/(1-b)q$ (identity)	1.20
b	Corporate borrowing per unit of capital	.30
d	Ratio of dividends to net earnings	.43
m	Firms' real gross marginal return on capital	(5)
q	Price of a share	1
R	Monthly nominal return on bonds (4–6 month commercial paper rate)	.00625
R_π	Change in R for 1 percent increase in π	1.0
r_b	Expected net real monthly return on bonds	−.00101
r_s	Expected net real monthly return on stocks	.00500
s_1	$[a(b-\lambda)-\theta_c]\sigma_1$ (identity)	-4.3×10^{-6}
s_2	$a(1-\tau_e)\sigma_2$ (identity)	.039
α	Expected real gross marginal return on capital (monthly)	.0086
β	Fraction of wealth in stocks	.67
γ	Index of relative risk aversion	
λ	Reduction in net profits per unit of capital from 1 percent increase in price level	.26
π	Expected monthly rate of inflation	.00539
ρ	Correlation coefficient: $E(dz_1\,dz_2)/dt$	−.22
σ_1^2	Variance of monthly inflation rate	4.6×10^{-6}
σ_2^2	Monthly variance of real gross marginal return on capital	.0030
Σ_{12}^2	$(1/dt)\mathrm{Var}(\xi_s-\xi_b)$	.0015
$\sigma_1\Sigma$	$(1/dt)\mathrm{Var}(\xi_b)-(1/dt)\mathrm{Cov}(\xi_s,\xi_b)$	-1.40×10^{-5}
ψ_s	Firms' real net rate of earnings per dollar of equity	(7)
ψ_s^n	Firms' nominal net rate of earnings per dollar of equity	(18)
ξ_b, ξ_s	Investors' real net rates of return on bonds and equity	(4), (10)
θ	Personal tax rate on interest and dividends	.30
θ_c	Effective tax rate on capital gains	.05
τ_s	Statutory corporate income tax rate	.48
τ_e	Effective corporate income tax rate	.40

[a] Or equation number.

and note that it must satisfy

$$\text{(A2)}\quad \delta V=\max_{C,\beta}\Big\{C^{1-\gamma}/(1-\gamma)$$
$$+\left[\beta(r_s-r_b)W+r_bW-C\right]V_W$$
$$+\left(\frac{1}{2}\beta^2\Sigma_{12}^2+\frac{1}{2}\sigma_1^2-\beta\sigma_1\Sigma\right)W^2V_{WW}\Big\}$$

where Σ_{12}^2 and Σ are defined in equations (25) and (26). The first-order conditions are

$$\text{(A3)}\quad C^*=V_W^{-1/\gamma}$$

and

$$\text{(A4)}\quad \beta^*=-(r_s-r_b)V_W$$
$$/\Sigma_{12}^2WV_{WW}+\sigma_1\Sigma/\Sigma_{12}^2.$$

Now substitute (A3) and (A4) into (A2) and solve the resulting differential equation for V to yield

$$\text{(A5)}\quad V(W)=A^{-\gamma}W^{1-\gamma}/(1-\gamma),$$

where

$$\text{(A6)}\quad A=\frac{1-\gamma}{\gamma}\left[\frac{\delta}{1-\gamma}-r_b\right.$$
$$-(r_s-r_b)^2/2\Sigma_{12}^2\gamma-(r_s-r_b)\sigma_1\Sigma/\Sigma_{12}^2$$
$$\left.+\frac{1}{2}\gamma\sigma_1^2\left(1-\Sigma^2/\Sigma_{12}^2\right)\right].$$

TABLE A2—COMPUTATION OF "ZERO-INFLATION" EFFECTIVE TAX RATE τ_e (Nonfinancial Corporations)

	Profits (1)	Interest (2)	Gross Profits (3)	Tax Liability (4)	Excess Tax due to Inflation (5)	Effective Taxes (6)	Effective Tax Rate (7)
1962	45.6	4.5	50.1	20.6	2.4	20.5	.41
1963	51.2	4.8	56.0	22.8	2.1	23.2	.41
1964	57.7	5.3	63.0	24.0	2.0	24.7	.39
1965	67.7	6.1	73.8	27.2	1.8	28.3	.38
1966	72.2	7.4	79.6	29.5	2.0	31.1	.39
1967	68.8	8.7	77.5	27.7	2.4	29.5	.38
1968	73.3	10.1	83.4	33.4	3.2	35.5	.43
1969	67.5	13.1	80.6	33.1	4.1	35.9	.45
1970	52.7	17.0	69.7	27.0	4.8	30.6	.44
1971	62.1	18.0	80.1	29.8	5.5	32.9	.41
1972	72.7	19.1	91.8	33.6	6.0	36.8	.40
1973	78.6	23.0	101.6	40.0	6.9	44.1	.43
1974	63.6	29.6	93.2	42.0	10.3	45.9	.49
1975	86.1	30.8	116.9	41.2	15.4	40.6	.35
1976	107.3	29.5	136.8	52.6	17.3	49.5	.36
1977	126.3	33.2	159.5	59.4	19.1	56.2	.35

Sources: Cols. (1), (2), (4): *Survey of Current Business* (September 1981, p. 50, Table (1.13); 1977 line: July 1981, p. 10); col. (5): Feldstein and Summers (1979).

Notes: Col. (1) is Profits with *IVA*, *CCA*; col. (2) is Interest (RbK); col. (3) is (αK), cols. (1)+(2); col. (6) is cols. (4)−(5)+ τ_s × col. (2); col. (7) is τ_e = col. (6)/col. (3).

Substituting (A5) into (A3) and (A4) gives $C^*(W)=AW$, and equation (24) for β^*.

C. *Calculating Changes in Share Values*

Based on the partial equilibrium relationship (28), one can calculate how the share price q changes in response to changes in π, σ_1^2, α, σ_2^2, or γ, as in Table 1. Note that β^* is given by equation (24), Σ_{12}^2 and Σ by equations (25) and (26), and r_b, r_s, s_1, and s_2 by equations (4) and (10). The relevant derivatives are

$$\text{(A7)}\quad d\log q/d\pi = B\big[a(b-\lambda) - a(1-\tau_s)bR_\pi + (1-\theta_c) - (1-\theta)R_\pi\big]/\gamma\Sigma_{12}^2$$

$$\text{(A8)}\quad d\log q/d\sigma_1^2 = B\big[\eta(1-\beta a(1-\tau_e)\rho\sigma_2/\sigma_1) - 1/\gamma + a(1-\tau_e)\rho\sigma_2/2\sigma_1 - \beta\eta^2\big]/\Sigma_{12}^2$$

$$\text{(A9)}\quad d\log q/d\alpha = Ba(1-\tau_e)/\gamma\Sigma_{12}^2$$

$$\text{(A10)}\quad d\log q/d\sigma_2^2 = B\big[a(1-\tau_e)\rho\sigma_1/2\sigma_2 - \beta a^2(1-\tau_e)^2 - \beta a(1-\tau_e)\rho\eta\sigma_1/\sigma_2\big]/\Sigma_{12}^2$$

$$\text{(A11)}\quad d\log q/d\gamma = -B(r_s-r_b)/\gamma^2\Sigma_{12}^2$$

where $\eta = a(b-\lambda)+1-\theta_c$, and

$$\text{(A12)}\quad B = q\gamma(1-b)\Sigma_{12}^2\Big/\Big\{q\beta\gamma(1-b)\Sigma_{12}^2 + (1-\theta)d\big[(1-\tau_e)\alpha-(1-\tau_s)bR + (b-\lambda)(\pi+\gamma\sigma_1^2)+(1-\tau_e)\gamma\rho\sigma_1\sigma_2 - (1-\tau_e)^2 2\beta a\sigma_2^2 - 2\beta(a(b-\lambda)+\eta) \times (1-\tau_e)\rho\sigma_1\sigma_2 + (b-\lambda)\eta\sigma_1^2\big]\Big\}.$$

REFERENCES

Black, Fischer, "Studies of Stock Price Volatility Changes," *Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economic Statistics Section*, 1976, 177–81.

Bodie, Zvi, "Common Stocks as a Hedge Against Inflation," *Journal of Finance*, May 1976, *31*, 459–70.

______, **Kane, Alex and McDonald, Robert,** "Inflation and the Role of Bonds in Investor Portfolios," Working Paper No. 1091, National Bureau of Economic Research, March 1983.

Bulow, Jeremy I., and Summers, Lawrence H., "The Taxation of Risky Assets," Working Paper No. 897, National Bureau of Economic Research, June 1982.

Fama, Eugene F., "Short-Term Interest Rates as Predictors of Inflation," *American Economic Review*, June 1975, *65*, 269–82.

______, "Stock Returns, Real Activity, Inflation, and Money," *American Economic Review*, September 1981, *71*, 545–65.

______ **and Schwert, G. William,** "Asset Returns and Inflation," *Journal of Financial Economics*, November 1977, *5*, 115–46.

Feldstein, Martin, (1980a) "Inflation and the Stock Market," *American Economic Review*, December 1980, *70*, 839–47.

______, (1980b) "Inflation, Tax Rules, and the Stock Market," *Journal of Monetary Economics*, July 1980, *6*, 309–31.

______, "Inflation and the Stock Market: Reply," *American Economic Review*, March 1982, *72*, 243–46.

______ **and Summers, Lawrence,** "Is the Rate of Profit Falling?," *Brookings Papers on Economic Activity*, 1:1977, 211–28.

______ **and** ______, "Inflation and the Taxation of Capital Income in the Corporate Sector," *National Tax Journal*, December 1979, *32*, 445–70.

______, **Poterba, James and Dicks-Mireaux, Louis,** "The Effective Tax Rate and the Pretax Rate of Return," Working Paper No. 740, National Bureau of Economic Research, August 1981.

Fischer, Stanley, "The Demand for Index Bonds," *Journal of Political Economy*, June 1975, *83*, 509–34.

Friedman, Milton, "Nobel Lecture: Inflation and Unemployment," *Journal of Political Economy*, June 1977, *85*, 451–72.

Friend, Irwin, and Blume, Marshall E., "The Demand for Risky Assets," *American Economic Review*, December 1975, *65*, 900–23.

______ **and Hasbrouck, Joel,** (1982a) "Inflation and the Stock Market: Comment," *American Economic Review*, March 1982, *72*, 237–42.

______ **and** ______, (1982b) "Effect of Inflation on the Profitability and Valuation of U.S. Corporations" in M. Sarnat and G. Szego, eds., *Savings, Investment, and Capital Markets in an Inflationary Economy*, Cambridge: Ballinger, 1982.

Grossman, Sanford J., and Shiller, Robert J., "The Determinants of the Variability of Stock Market Prices," *American Economic Review Proceedings*, May 1981, *71*, 222–27.

Hendershott, Patric H., "The Decline in Aggregate Share Values: Taxation, Valuation Errors, Risk, and Profitability," *American Economic Review*, December 1981, *71*, 909–22.

Holland, Daniel M., and Myers, Stewart C., "Profitability and Capital Costs for Manufacturing Corporations and All Nonfinancial Corporations," *American Economic Review Proceedings*, May 1980, *70*, 320–25.

Malkiel, Burton G., "The Capital Formation Problem in the United States," *Journal of Finance*, May 1979, *34*, 291–306.

Merton, Robert C., "Optimum Consumption and Portfolio Rules in a Continuous-Time Model," *Journal of Economic Theory*, December 1971, *3*, 373–413.

______, "On Estimating the Expected Return on the Market," *Journal of Financial Economics*, December 1980, *8*, 323–61.

Modigliani, Franco, and Cohn, Richard, "Inflation, Rational Valuation, and the Market," *Financial Analysts Journal*, March 1979, *35*, 3–23.

Nelson, Charles R., "Inflation and Rates of Return on Common Stocks," *Journal of Finance*, May 1976, *31*, 471–83.

______ **and Schwert, G. William,** "On Testing the Hypothesis that the Real Rate of Interest is Constant," *American Economic Review*, June 1977, *67*, 478–86.

Parks, Richard W., "Inflation and Relative Price Variability," *Journal of Political Economy*, February 1978, *86*, 79–95.

Shiller, Robert J., "Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends?," *American Economic Review*, June 1981, *71*, 421–36.

Summers, Lawrence H., (1981a) "Inflation, the

Stock Market, and Owner-Occupied Housing," *American Economic Review Proceedings*, May 1981, *71*, 429–34.

______, (1981b) "Inflation and the Valuation of Corporate Equities," Working Paper No. 824, National Bureau of Economic Research, December 1981.

______, "The Non-Adjustment of Nominal Interest Rates: A Study of the Fisher Effect," in J. Tobin, ed., *Macroeconomics, Prices & Quantities*, Washington: The Brookings Institution, 1983.

U.S. Department of Commerce, *National Income and Product Accounts of the United States, 1929–79*, *Survey of Current Business*, Suppl. July 1981; September 1981.

Welfare Costs per Dollar of Additional Tax Revenue in the United States

By CHARLES STUART*

Assessing the optimal level of government spending is as difficult as it is important. On a theoretical level, the issue can be stated simply as one of comparing the marginal benefits and costs of public expenditures. It is the measurement of the benefits and costs that presents problems. In this paper, I present fairly striking calculations of the costs of marginal governmental expenditures.

An insight by Edgar Browning serves as the starting point for the analysis. Browning (1976) observed that the social cost of financing a marginal dollar of public expenditure is the sum of that dollar, which is diverted from private use, plus the change in the total welfare cost of taxation caused by increasing tax revenue by the dollar. This latter component will be termed "marginal excess burden" in what follows. It can be regarded as a per dollar surcharge that must be borne whenever the public sector alters the allocation or distribution of resources through fiscal measures. Not surprisingly, the notion of marginal excess burden plays a central role in theories of optimal taxation. For instance, Peter Diamond and Daniel McFadden (1974, p. 12) point out that optimal taxation requires equality of marginal excess burdens across revenue sources. Similarly, Dan Usher (1982) reworks the analysis of Anthony Atkinson and Nicholas Stern (1974) to argue that Browning's marginal cost of public funds (one plus marginal excess burden) enters into the first-order conditions for the optimal provision of publicly supplied goods.

Using a partial-equilibrium approach based on Arnold Harberger's (1964) excess burden formula, Browning calculated that the value of the marginal excess burden from labor income in the United States was on the order of 9¢ to 16¢ on the dollar in 1974. This would mean that a dollar of public funds was efficiently spent only if it generated social benefits of at least $1.09 to $1.16. However, there are a number of reasons for questioning Browning's partial-equilibrium treatment. First, the Harberger formula is exact only in the neighborhood of an undistorted equilibrium for an economy with a linear production frontier (Harberger, 1971, p. 792). Here, the undistorted equilibrium requirement accords poorly with today's significant marginal tax rates on labor income while the linearity condition seems implausible in light of the literature on the magnitude of the elasticity of substitution in the aggregate production function. Second, and possibly more importantly, the Harberger formula is conceptually inadequate for measuring marginal excess burden. It is certainly true that this formula correctly measures the cost of failing to use lump sum taxation. However, lump sum taxation is not the alternative foregone in raising an additional dollar of tax revenue. To calculate the welfare cost of raising an additional dollar of revenue, one wishes to compare changes in utility and revenue as the economy moves from an equilibrium before a tax increase to one after the increase. The Harberger formula does not do this. Instead, it compares an undistorted equilibrium to a hypothetical, fully compensated allocation (Diamond and McFadden). The problem is that the change in the level of tax revenue as the economy moves between fully compensated points does not generally equal the change in revenue that the actual (uncompensated) economy experiences, so estimates of marginal

*Department of Economics, University of California, Santa Barbara, CA 93106, and University of Lund, Sweden. This paper was formerly titled "Marginal Excess Burden in the U.S." I am indebted to Ingemar Hansson for numerous valuable comments on different versions of this work. Thanks are also due to Edgar Browning, Don Fullerton, and John Riley. The research support of the Bank of Sweden Tercentenary Foundation is gratefully acknowledged.

excess burden based on the Harberger formula are biased.[1] A related problem is that since the equilibrium level of tax revenue generally depends on the way in which the government spends the revenue, the value of marginal excess burden cannot itself be independent of the type of marginal spending (compare Atkinson and Stern, equation (3); my 1982 paper). In the Harberger-Browning approach, however, the dependence of marginal excess burden on the use of marginal public revenue via the government budget constraint fails to show up. A third difficulty with Browning's calculation is that while he computes welfare loss triangles arising when taxation reduces the amount of labor supplied, he assumes that taxation does not reduce the tax base (compare his equations (4) and (5)).

All of these difficulties can be overcome by estimating marginal excess burden in a simple, general-equilibrium framework. Further, such an approach can provide valuable information on how marginal excess burden varies with the type of government spending or with other policy or structural parameters. In this paper, I adapt the model developed in my earlier paper (1981) to calculate the marginal excess burden from taxes on labor income in the United States. For simplicity, I follow Browning's assumptions—especially on labor supply elasticities—as closely as possible. To highlight the potential magnitude of marginal excess burden, however, estimates based on higher but not implausible labor supply elasticities are also reported.

In Section I, the structure of the model used for the calculations is presented. Parameterization corresponding to U.S. economic experience is treated in Section II. Section III contains the results as well as sensitivity tests.

I. The Model

Briefly, the model contains a single, utility-maximizing, aggregate household that allocates a fixed amount of labor time between a taxed and an untaxed sector. Each sector is represented by a production function, and each has a fixed and immobile capital stock. Revenue from the taxation of labor in the taxed sector is partially redistributed to the household as a lump sum and partially expended on government consumption. The government budget balances. An increase in the marginal tax rate causes labor to flow from the taxed to the untaxed sector and thus influences the equilibrium consumptions of taxed and untaxed sector outputs. Consequent welfare changes are evaluated by calculating the compensation required to equate the utility levels of the pre-tax-increase and the post-tax-increase consumption vectors.

Let L denote the household's total labor time, L_1 be the amount of time devoted to taxed uses (i.e., to the taxed sector), and L_2 be the amount allocated to untaxed uses (sector). Assume that

$$(1) \qquad L_1 + L_2 = L.$$

A rough picture of the dichotomization between L_1 and L_2 is that the former corresponds to normal market employment while the latter encompasses time devoted to home production and leisure, as well as to "on-the-job leisure" and to activities leading to tax evasion and fringe benefits.

The outputs produced by taxed and untaxed uses of labor, denoted Y_1 and Y_2, respectively, are represented by Cobb-Douglas production functions:

$$(2a) \qquad Y_1 = AL_1^a,$$

$$(2b) \qquad Y_2 = BL_2^b,$$

where A, a, B, and b are parameters.[2] Capital stocks in each sector are constant and are hence subsumed in A and B.[3] The

[1] Under certain conditions (for example, homotheticitiy and separability of publicly supplied goods in utility) the bias is upward (see my 1982 paper).

[2] Early cross-section work on the elasticity of substitution in the aggregate production function for the United States (Jora Minasian, 1961; Robert Solow, 1964; Frederick Bell, 1964) supports a Cobb-Douglas specification, at least for the taxed sector. Stronger support is in work by Zvi Griliches (1967).

[3] This assumption should cause downward-biased estimates of the welfare cost of tax increases—for dis-

parameters a and b represent labor's shares. Taxed-sector output is taken as the numeraire.

The household maximizes a Stone-Geary generalized *CES* utility function:

$$(3)\quad U=\left[\alpha\overline{Y}_1^{-\rho}+(1-\alpha)\left(\overline{Y}_2-\delta\right)^{-\rho}\right]^{-1/\rho},$$

where $\overline{Y}_1$ and $\overline{Y}_2$ are the amounts of taxed- and untaxed-sector consumption, respectively, and where α, ρ, and δ are parameters. The maximization is performed by allocating labor time to the two sectors in accord with (1) so as to influence the marginal consumptions of taxed- and untaxed-sector output.

Since no tax wedge exists in the untaxed sector, the amounts of sector-two output produced and consumed are equal:

$$(4)\qquad Y_2=\overline{Y}_2,$$

and hence the increase in the consumption of untaxed-sector output given a one unit increase in L_2 is simply

$$\partial\overline{Y}_2/\partial L_2=\partial Y_2/\partial L_2=bBL_2^{b-1}.$$

To represent the household's perceived increase in $\overline{Y}_1$ obtained by a small increase in L_1, assume that any redistribution of tax revenue to the household is regarded as a lump sum and that any capital income, $(1-a)Y_1$, is similarly viewed as a lump sum. Letting w be the gross wage and t' be the (constant) marginal tax rate, the household thus sees the net wage as the marginal determinant of $\overline{Y}_1$, that is,

$$\partial\overline{Y}_1/\partial L_1=w(1-t').$$

The first-order condition for utility maximization is then

$$(5)\quad (1-t')w=\left\{(1-\alpha)/\alpha\left[\left(\overline{Y}_2-\delta\right)/\overline{Y}_1\right]^{-\rho-1}\right\}bBL_2^{b-1},$$

or that the number of units of $\overline{Y}_1$ obtained on a non-lump-sum basis per marginal unit of L_1 equals the number of units of $\overline{Y}_2$ obtained per marginal unit of L_2 times the marginal rate of substitution (in braces).

The government budget is assumed to balance with tax revenue from labor income being expended on redistribution (R) and government consumption (G):

$$(6)\qquad R+G=twL_1,$$

where t is the average tax rate on labor income. To allow the model to treat situations where marginal tax revenue is expended only on R, or only on G (or on a mixture of the two), it is convenient to write G as a linear function of total revenue from labor income:[4]

$$(7)\qquad G=g_0+g_1(twL_1).$$

When $g_1=0$, the model then directs all marginal tax revenue to spending on R, while when $g_1=1$, all marginal revenue is assumed to be spent on G. Note that the initial level of G can be set to any feasible value by an appropriate choice of g_0.

With the public sector specified in this way, there is a linear relationship between taxed-sector output and consumption. Writing $\overline{Y}_1$ as the sum of net wages, capital income, and redistributions, this relationship is

$$(8)\quad \overline{Y}_1=(1-t)wL_1+(1-a)Y_1+R$$
$$=(1-g_1ta)Y_1-g_0,$$

cussion, see my 1981 article. As well, disaggregating to more than two sectors should increase welfare costs (see John Shoven, 1976).

[4] Government consumption does not explicitly enter the utility function (3). An interpretation is that the model treats government consumption as providing utility in a way which is mathematically separable from $\overline{Y}_1$ and $\overline{Y}_2$ (and which is notationally suppressed in (3)). Thus g_1 can be regarded as the share of marginal tax revenue that is publicly spent on goods that are separable from $\overline{Y}_1$ and $\overline{Y}_2$, and which therefore do not influence the allocation of labor to L_1 or L_2 (compare condition (5)). Note that "public good-like" benefits from redistributions (over and above the direct utility value of the Y_1 obtained by recipients) can also be introduced without disturbing the present structure and results by merely appending a separable function of R to equation (3).

where the final equality uses (2a), (6), and (7), as well as the assumption that the wage in the taxed sector is set competitively:

$$(9) \qquad w = aAL_1^{a-1}.$$

In order to close the model, it is necessary to specify how the average tax rate in the economy changes when the government raises the marginal rate. For simplicity, it is assumed that the ratio of the marginal to the average tax rate is a constant, τ:[5]

$$(10) \qquad t = t'/\tau.$$

For any initial value of the marginal tax rate, numerical solution of (1), (2a), (2b), and (4)–(10) yields equilibrium values of t, L_1, L_2, Y_1, Y_2, $\overline{Y}_1$, $\overline{Y}_2$, w, and the two uses of tax revenue, G and R. It now remains to calculate welfare effects. First, suppose a tax increase causes the household's equilibrium consumption to change from $(\overline{Y}_1, \overline{Y}_2)$ to $(\overline{Y}_1', \overline{Y}_2')$. Denote the numeraire value of the resulting reduction in household utility by ΔC. Then ΔC can be measured as the amount of taxed-sector output (the numeraire) that would just be needed to restore the household to its original utility level; that is, by the root of $U(\overline{Y}_1, \overline{Y}_2) = U(\overline{Y}_1' + \Delta C, \overline{Y}_2')$.[6] When all marginal tax revenue is expended on government consumption ($g_1 = 1$), this root includes changes in both the direct burden (i.e., in tax revenue) and in the excess burden of taxation, so the change in excess burden is $\Delta C - \Delta(twL_1)$. Since marginal excess burden is defined here as the change in excess burden *per additional dollar of tax revenue*, the correct expression for marginal excess burden is therefore $\Delta C/\Delta(twL_1) - 1$ when $g_1 = 1$. On the other hand, when all tax revenue is returned to taxpayers on the margin ($g_1 = 0$), the direct burden is balanced by a change in redistributions and is thus in effect netted out of ΔC. Marginal excess burden is then conveniently taken as simply $\Delta C/\Delta(twL_1)$. In general, netting redistributions out of ΔC to find excess burden entails calculating marginal excess burden as[7]

$$(11) \qquad MEB = \Delta C/\Delta(twL_1) - g_1.$$

II. Parameterization

This section describes how values for A, a, B, b, L, α, δ, ρ, g_0, g_1, and τ were chosen so as to be consistent with economic experience in the United States. Whenever possible, parameterization was undertaken using data from 1976, which was a year of "normal" business activity, being neither a business cycle peak nor a trough.

Some of the parameters can be chosen with more certainty than others. These will be discussed first.

A. *Tax Rates*

The marginal tax rate in 1976 is taken as the weighted average of marginal rates for different brackets, with weights equal to shares of income in each bracket.[8] Income,

[5] It should be emphasized that this relationship reflects the way in which the government changes the tax schedule over time to raise or lower total tax revenue and *not* just the shape of the tax schedule at any point in time. For instance, one may think of the government beginning with a linear tax schedule, raising the slope of the schedule (t') incrementally, and then adjusting the intercept. Equation (10) merely specifies that the intercept is to be adjusted so that the average tax rate in the new equilibrium exceeds the initial average tax rate by exactly $1/\tau$ times the increment in t'. The assumed relationship is adequate to closely approximate the tax-change assumptions made by Browning in two of his three cases (i.e., proportional and "degressive" taxes).

[6] This is John Hicks' (1946) "compensating surplus." A justification for using it is in my 1982 paper. Alternative welfare measures that might instead be used are the equivalent surplus, the compensating variation, and the equivalent variation. In the implementation of the model described here, substitution of these alternative measures was found to affect the results only at the third significant digit.

[7] This formulation neglects administration and other compliance costs of running the tax system. Browning (p. 293) cites a figure of 2–2.5 percent of tax revenues for these costs.

[8] Browning calculates changes in excess burdens for taxpayers in different income classes and sums to get the aggregate change in excess burden. His approach is equivalent for proportional or degressive tax systems to finding the marginal excess burden for an aggregate household whose tax rate is a weighted average of the marginal rates for each income class, with weights equal to class income shares.

TABLE 1—GOVERNMENT SPENDING, 1976[a]

Counted as Government Consumption		Counted as Redistributions	
Federal		**Federal**	
National Defense	89.4	Agriculture	2.5
International Affairs	5.6	Commerce and Housing Credit	3.8
Natural Resources and Development	8.1	Education, Training, Employment and, Social Services	18.7
Community and Regional Development	4.8	Health	33.4
Administration of Justice	3.3	Income Security	127.4
General Government	3.0	Veterans' Benefits	18.4
(1/2) General Space, Science, and Technology	2.2	(1/2) General Space, Science, and Technology	2.2
(1/2) Energy	1.6	(1/2) Energy	1.6
(1/4) Transportation	3.4	(3/4) Transportation	10.1
State and Local		**State and Local**	
Highways	23.9	Education	97.2
(4/5) Other	82.4	Public Welfare	32.6
		(1/5) Other	20.6
Total	$G = 227.7$		$R = 368.5$

[a]Amounts are in billions of 1976 dollars.

payroll, and indirect taxes as well as the tax effect of income-indexed transfers are included since all of these can be avoided if labor is shifted from taxed to untaxed uses. Data on federal marginal rates by brackets are published by the Internal Revenue Service (1979). On the basis of work by Browning and William Johnson (1979, pp. 63–65), the progressivities of federal, and state and local income taxes are assumed to be equal. Data on the tax effect of transfer payments are also from Browning and Johnson. Calculated values for 1976 are $t' = .427$ and $t = .273$. Dividing the former by the latter yields the estimate $\tau = 1.564$. This is close to the value implicit in Browning's "degressive tax" case (1.629). (Using Browning's figure instead of 1.564 would increase marginal excess burden slightly.)

B. *Government Consumption*

The model sketched above dichotomizes government expenditures into government consumption, which is assumed to have no influence on the marginal rate of substitution between the outputs of the two sectors, and redistributions to the household, which are treated as perfect substitutes for private consumption of taxed-sector output. The procedure for estimating g_0 and g_1 was to interpret federal, and state and local budget outlays (*Economic Report of the President*, 1981, Tables B-70, B-76) as either G or R for 1976 and 1971. The assumptions are detailed in Table 1; these implied 1976 values of $G = 227.7$ and $R = 368.5$.[9] Similar figures for 1971, expressed in 1976 dollars (using the *CPI*) were $G = 235.1$ and $R = 290.5$. The recent historical pattern thus appears to be that government consumption has been relatively constant while redistributions have increased. Parameter values of $g_1 = 0$ and $g_0 = 227.7$ ($= G$ in 1976) reflect this. Most simulations are run with these values. However, several simulations are also run on the assumption that all marginal tax revenue is spent on G instead of R. To capture this

[9] This procedure is not as exact as one might like. For instance, roughly 20 percent of state and local revenues are provided by the federal government, so some expenditures may be double counted. This makes the assumed 1976 value of G too high and imparts a slight conservative bias to the results. A possible countervailing effect is that some of the assumed components of R may in reality be less than perfect substitutes for private income.

given $G = 227.7$ involves assuming that $g_1 = 1$ and $g_0 = -63.1$.

C. *Labor*

Not all of the 24 hours in a day are subject to an allocational choice between L_1 and L_2. It is assumed that 10 hours per day can be freely allocated (sensitivity analysis indicates that this assumption is innocuous, see Section III). On a yearly basis, this is 3,660 hours, which is taken as the value of L. In 1976, there were 157.32 billion hours of work in the taxed portion of the U.S. economy.[10] During the same year, the noninstitutional population 16 years of age and older was 156.05 million (*Economic Report of the President*, p. 264). Thus the average hours of work per employable person in 1976 was 1008.14. The 1976 allocation of labor is therefore taken to be $L_1 = 1008.14$ and $L_2 = 2651.86$.

D. *Taxed-Sector Production*

Net national product for 1976 was 1527.4 billion dollars (*Survey of Current Business*, 1977, p. 8). This is taken to be the value of Y_1. (Thus $\overline{Y}_1 = 1527.4 - 227.7 = 1299.7$ in 1976.) Compensation of employees was 1036.3 and proprietors' incomes were 88.0 (*SCB*, p. 9) If it is assumed that labor's share of Y_1 equaled labor's share of proprietors' incomes then total labor income is $1036.3 + (88.0)\ a$. This, with $a =$ total labor income/1527.4 yields $a = .720$. Plugging the assumed values of Y_1, L_1, and a into (2a) gives $A = 10.506$.

E. *Untaxed-Sector Production*

Little empirical analysis has been directed at determining the shape of the untaxed-sector production function. On the basis of the observed stability of the Cobb-Douglas function (see Paul Douglas, 1976), however, the naive assumption that $b = a = .720$ is not unreasonable (sensitivity analysis suggests that the choice of b is unimportant, see Section III). The value of B merely determines the units in which Y_2 is measured and has no influence on the results. A useful choice is to pick B so that the marginal rate of substitution in consumption equaled one as of 1976. Inserting (9) into (5) to eliminate w, imposing $MRS = 1$, and setting t', a, A, L_1, L_2, and b to their assumed (1976) values then implies $B = 7.892$.

F. *Utility*

Data on labor supply elasticities are used to fix α, δ, and ρ. In particular, Browning's survey of the (early) labor supply literature led him to assume that the compensated supply elasticity for labor was .2. From the studies referenced by Browning, a reasonable estimate of the uncompensated elasticity might be zero (i.e., elasticities for males were negative while female elasticities were generally positive and somewhat greater in magnitude). I therefore derive the compensated and uncompensated labor supply elasticities implied when the household's first-order condition (5) is differentiated first with respect to the gross wage and then with respect to lump sum income, and impose the constraints that these elasticities equal .2 and zero, respectively. A third constraint (from above) is that $MRS = 1$ given the 1976 values of all variables. Solution of these three relationships yields $\alpha = .9429$, $\delta = 1968.36$, and $\rho = 1.0625$. Deriving parameters in this way "calibrates" the model so that its equilibrium solution exactly replicates the 1976 magnitudes of the endogenous variables when the marginal tax rate t' is set to its 1976 level.

III. Results

Given the set of parameters generated by the procedure described in Section II, simulations were run by letting the marginal tax rate increase in one-percentage-point increments. The results are labeled "benchmark case" in Table 2. Note that the assumptions

[10] From *Survey of Current Business* (1977, p. S15), 151.39 billion hours were worked by waged and salaried employees in nonagricultural establishments in 1976. At the same time, 84,188 (000's) persons were employed in nonagricultural industries and 3,297 (000's) were employed in agriculture (*SCB*, p. S13). The text figure of $157.32 = (151.39)(1 + 3297/84188)$.

TABLE 2—RESULTS

	Marginal Excess Burden at		
Explanation	$t'=.427$	$t'=.46$	t^*
0. Benchmark Case	.207	.244	85
1. Spending on Government Consumption	.072	.090	88
2. $b=.95$	.207	.245	85
3. $L=5490$ hours/year	.207	.245	85
4. No Payroll Tax	.234[a]	.278[b]	84
5. $\eta=.318$	.574	.719	71
6. $\eta=.318$, Spending on Government Consumption	.427	.533	72
7. $\eta=.636$	.999	1.33	63

Note: Marginal excess burdens reported here contain no adjustments for administrative costs of collecting taxes. Unless otherwise stated, marginal public spending is on redistributions; η is the aggregate uncompensated wage elasticity of labor supply; t^* is shown in percent.

[a] $t'=.377$.

[b] $t'=.41$.

in this case correspond closely to those made by Browning in his analysis of degressive taxes. At a 41.3 percent marginal tax rate, Browning estimated marginal excess burden to be 13.4¢. The figure for the benchmark here is 20.7¢ of welfare loss per additional dollar of tax revenue at the 42.7 percent marginal tax rate that prevailed in 1976. This is roughly 1.5 times Browning's estimate. Note that if taxes have increased from 1976 to the present, marginal excess burden under this set of assumptions would be still higher today. It is difficult to know just what has happened to the aggregate marginal tax rate in the years since 1976. A rough estimate derived from data in the 1981 *Economic Report of the President* is as follows. Taking the sum of federal receipts from the individual income tax, Social Security tax, excise taxes, and customs duties (Table B-70), together with state and local revenues from sales and individual income taxes (Table B-76), and dividing by net national product (Table B-17) yields 0.210 for 1976 and 0.230 for 1979. This suggests that the average tax rate on L_1 may have risen by about two percentage points from 1976 to 1979. Given the value of τ assumed in Section II, the marginal tax rate on L_1 may thus be approximately 46 percent today. This would imply a marginal excess burden of 24.4¢.[11]

Also revealed in Table 2 is the marginal tax rate at which simulated tax revenue peaks. This rate is denoted t^* and equals 85 percent in the benchmark. Note that a similar result has been obtained by Don Fullerton (1982); in a more detailed model of the U.S. economy, he calculated that total tax revenue would reach a maximum at $t^*=.75$. By the way it is defined, marginal excess burden becomes infinite at t^* and is not a particularly useful concept beyond t^*.

Seven alternative sets of parameterizing assumptions were also simulated to examine the sensitivity of the results. The effects of these alternative assumptions are also displayed in Table 2. The basic method for finding parameter values is unchanged. By the numbers:

1) *Spending on government consumption.* In the benchmark, all marginal tax revenue was redistributed on a lump sum basis. This assumption approximates the historical trend more closely than would the alternative specification that spending was on government consumption. When one sets $g_1=1$ (and $g_0=-63.1$) so that all marginal tax revenue is funneled into government consumption, the calculated 1976 value of marginal excess burden drops to 7.2¢. The

[11] A caveat is that the net impact on t' of developments since 1979—including bracket creep, payroll tax increases, and the 1981 tax cut—is uncertain. In any case, comparing marginal excess burden at $t'=.427$ and at $t'=.46$ provides a feel for the sensitivity of the results to the level of t'.

size of this reduction—from double to roughly one-half of Browning's estimated value—provides strong confirmation that the ultimate use of public funds matters. There is an interesting and important explanation for the decline. Redistribution of tax revenue to taxpayers induces an income effect that increases the tendency for labor to leave the taxed sector when tax rates rise. This makes tax revenue increase less rapidly than would be the case if public spending were directed toward government consumption. The net effect is to reduce the denominator in equation (11); that is, to make the change in excess burden per dollar of additional revenue greater. A striking implication is that the relevant marginal excess burden for national defense is likely to be lower than the marginal excess burden for a redistributional social program. In very much the same way, the relevant marginal excess burden for wasteful government programs (i.e., programs that use resources but produce nothing of value) is lower than the marginal excess burden for redistributional social programs.

2) *Sensitivity analysis, b.* In parameterizing the untaxed-sector production function, the *ad hoc* assumption that $b = a = .720$ was employed. Here, the assumption that $b = .95$ is used instead. The effect of this replacement on the results is nil.

3) *Sensitivity analysis, L.* The benchmark parameterization assumed that 10 hours per day, or 3,660 hours in 1976, could be freely allocated to L_1 and L_2. The sensitivity of the results to this assumption is assessed by specifying instead that 15 hours per day can be freely allocated to L_1 and L_2. Thus $L = 5490$ (hours/year). Again, there is no effect on the results. The explanation is that the assumed wage and income elasticities of labor supply are the critical determinants of marginal excess burden. Thus with these elasticities held constant, changes in L (or b) induce compensating adjustments in the parameters of the utility function and no change in marginal excess burden occurs.

4) *Sensitivity analysis, payroll tax not treated as a tax.* One might argue that part of the Social Security payroll tax reflects a forced payment for individual retirement benefits that would be purchased voluntarily even in the absence of the Social Security system. Such a view implies that part of the payroll tax is not distortionary and hence should not, for purposes of welfare analysis, be included in t' and t. To assess this, all payroll taxes are netted from t' and t in the present scenario and the model is reparameterized at the reduced 1976 tax rates. Netting out payroll taxes lowers the 1976 marginal tax rate from .427 to .377. The average rate declines more sharply, falling from .273 to .174. As a consequence, the implied value of τ increases; that is, eliminating the (regressive) payroll tax causes the (remaining) tax structure to become more progressive. Simulation then yields a 1976 marginal excess burden (at $t' = .377$) of .234, which amounts to a small rise from the benchmark case. This rise is largely due to the implied increase in τ; without it, marginal excess burden would fall since assumed tax rates are lower.

5) *Higher assumed labor supply elasticity.* The labor supply elasticities assumed above are low by the standards of the recent literature. For instance, Fullerton examines male and female elasticities as cited in a survey by Mark Killingsworth (1983), weights these by relative income shares, and concludes that the aggregate uncompensated labor supply elasticity is +.15. Even this aggregate value may be too low as it is based largely on a relatively older and less sophisticated body of studies. In particular, recent work on male labor supply tends to produce positive elasticity estimates as often as negative ones, although absolute magnitudes are generally small.[12] Two important examples are studies by Thomas MaCurdy (1981), who considers labor supply in a life cycle setting and obtains elasticites in the range .05 to .13, and by B. K. Atrosic (1982), who considers variations in preferences in a model with flexible functional forms and finds male elasticities of .19 to .39. My reading is therefore that a

[12] Studies by Julie DaVanzo, Dennis De Tray, and David Greenberg (1976) and by George Borjas (1980) have examined several questionable practices common in earlier work (i.e., regarding wage and asset values as exogenous, severely restricting samples, and defining the independent wage variable as income divided by the dependent variable, hours worked). Simple corrections for these practices generally caused elasticities to change in sign from negative to positive.

TABLE 3—RECENT ESTIMATES OF THE WAGE ELASTICITY OF FEMALE LABOR SUPPLY

Study	Elasticity	Notes
Rosen (1976)	2.30	Corrects for nonlinearities in budget due to taxes; uses Tobit to deal with observations with zero supplied labor.
Heckman (1976)	4.31	Corrects for sample-selection bias.
Cogan (1980)	2.45	Corrects for selection bias; allows fixed costs of working. Estimate is from Heckman et al. (1981), which refers to this study as Cogan (1980).
Schultz (1980)	1.26	Uses Tobit. Reported figure is average elasticity (over age cohorts) for whites; average for blacks is 0.88.
Heckman (1980)	4.83	Corrects for selection bias; allows fixed costs of working; treats labor market experience as endogenous.
Hanoch (1980)	1.44	Corrects for selection bias; allows fixed costs of working; allows simultaneous determination of annual hours and weeks worked; treats 52 weeks/year of work as corner solution.
Hausman (1981)	0.91	Corrects for nonlinear budget caused by taxes and income-indexed transfers. Reported elasticity is evaluated at means for women who work.

zero uncompensated wage elasticity for males is not an unreasonable assumption.

For females, a survey by James Heckman et al. (1981) partitions the literature into "first-generation" and "second-generation" studies. Browning cities the simpler first-generation studies exclusively; according to Heckman et al., these studies report elasticities between −.1 and +1.6. The later, second generation work, on the other hand, attempts to correct for the presence of discontinuities in the labor supply function (due to fixed costs of working), nonlinear budget constraints (due to taxes), sample selection biases, and/or endogeneity of wage and asset variables. A digest of the more recent studies, with elasticity estimates, is in Table 3. The average elasticity in the table is 2.5. On this basis, I would think that an assumed female wage elasticity of 1.0 is not implausibly high; indeed, an elasticity of 2.0 is not completely out of the ballpark given the distribution of estimates of recent studies. Since the relative shares of labor income for males and females were .682 and .318, respectively, in 1976 (*Current Population Reports*, 1978, Table 49), these elasticity assumptions for females imply aggregate elasticities of .318 and .636, respectively, when combined with a zero elasticity for males. While the latter figure may be on the high side, the former is, again, not unreasonable given the recent evidence.

Accordingly, simulation 5 modifies the benchmark by assuming that the uncompensated wage elasticity is .318. As in the benchmark, the compensated elasticity is taken as the uncompensated elasticity plus .2. This causes marginal excess burden to rise to 57.4¢ on the dollar at the 1976 tax rate and to roughly 72¢ on the dollar at a 46 percent marginal tax rate on labor income.

6) *Higher elasticity with spending on government consumption.* In simulation 5, marginal public spending was redistributional; here, marginal spending is instead assumed to be on government consumption. As was the case in comparing simulations 1 and 2, government consumption entails a lower marginal excess burden (.427 at 1976 tax rates). Indeed, the absolute difference in marginal excess burden is roughly the same between the two pairs of simulations. That is, a shift from redistribution to government consumption lowers marginal excess burden by about 14¢.

7) *A still higher elasticity.* As a final sensitivity test, the implications of assuming the high labor supply elasticity of .636 are examined. The compensated elasticity is

taken to be .836. This causes calculated marginal excess burden to rise to one dollar or more per dollar of tax revenue for redistributional spending.

REFERENCES

Atkinson, Anthony, and Stern, Nicholas, "Pigou, Taxation and Public Goods," *Review of Economic Studies*, January 1974, *41*, 119–28.

Atrosic, B. K., "The Demand for Leisure and Nonpecuniary Job Characteristics," *American Economic Review*, June 1982, *72*, 428–40.

Bell, Frederick, "The Role of Capital-Labor Substitution in the Economic Adjustment of an Industry Across Regions," *Southern Economic Journal*, October 1964, *31*, 123–31.

Borjas, George, "The Relationship between Wages and Weekly Hours of Work: The Role of Division Bias," *Journal of Human Resources*, Summer 1980, *15*, 409–23.

Browning, Edgar, "The Marginal Cost of Public Funds," *Journal of Political Economy*, April 1976, *84*, 283–98.

______ **and Johnson, William,** *The Distribution of the Tax Burden*, Washington: American Enterprise Institute, 1979.

Cogan, John, "Labor Supply with Costs of Labor Market Entry," in James Smith, ed., *Female Labor Supply*, Princeton: Princeton University Press, 1980.

DaVanzo, Julie, De Tray, Dennis and Greenberg, David, "The Sensitivity of Male Labor Supply Estimates to Choice of Assumptions," *Review of Economics and Statistics*, August 1976, *55*, 313–25.

Diamond, Peter and McFadden, Daniel, "Some Uses of the Expenditure Function in Public Finance," *Journal of Public Economics*, February 1974, *3*, 3–21.

Douglas, Paul, "The Cobb-Douglas Production Function Once Again: Its History, Its Testing, and Some New Empirical Values," *Journal of Political Economy*, October 1976, *84*, 903–15.

Fullerton, Don, "On the Possibility of an Inverse Relationship Between Tax Rates and Government Revenues," *Journal of Public Economics*, October 1982, *19*, 3–22.

Griliches, Zvi, "Production Functions in Manufacturing: Some Preliminary Results," in M. Brown, ed., *The Theory and Empirical Analysis of Production*, Studies in Income and Wealth, Vol. 31, New York: Columbia University Press, 1967, 275–322.

Hanoch, Giora, "A Multivariate Model of Labor Supply: Methodology and Estimation," in James Smith, ed., *Female Labor Supply*, Princeton: Princeton University Press, 1980.

Harberger, Arnold, "Taxation, Resource Allocation, and Welfare," in *The Role of Direct and Indirect Taxes in the Federal Reserve System*, Princeton: Princeton University Press, 1964.

______, "Three Basic Postulates for Applied Welfare Economics," *Journal of Economic Literature*, September 1971, *9*, 785–97.

Hausman, Jerry, "Labor Supply," in H. Aaron and J. Pechman, eds., *How Taxes Affect Economic Behavior*, Washington: The Brookings Institution, 1981.

Heckman, James, "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement*, December 1976, *5*, 475–92.

______, "Sample Selection Bias as a Specification Error with an Application to the Estimation of Labor Supply Functions," in James Smith, ed., *Female Labor Supply*, Princeton: Princeton University Press, 1980.

______, **Killingsworth, Mark and MaCurdy, Thomas,** "Empirical Evidence on Static Labour Supply Models: A Survey of Recent Developments," in Z. Hornstein, et al., eds., *The Economics of the Labour Market*, London: HMSO, 1981.

Hicks, John R., "The Generalized Theory of Consumer's Surplus," *Review of Economic Studies*, No. 2, 1946, *13*, 68–74.

______, *A Revision of Demand Theory*, Oxford: Clarendon, 1956.

Killingsworth, Mark, *Labor Supply*, New York: Cambridge University Press, 1983.

MaCurdy, Thomas, "An Empirical Model of Labor Supply in a Life-Cycle Setting," *Journal of Political Economy*, December 1981, *89*, 1059–85.

Minasian, Jora, "Elasticities of Substitution and Constant-Output Demand Curves for Labor," *Journal of Political Economy*, June 1961, *69*, 261–70.

Rosen, Harvey, "Taxes in a Labor Supply Model with Joint Wage-Hours Determination," *Econometrica*, May 1976, *44*, 485–507.

Schültz, T. Paul, "Estimating Labor Supply Functions for Married Women," in James Smith, ed., *Female Labor Supply*, Princeton: Princeton University Press, 1980.

Shoven, John, "The Incidence and Efficiency of Taxes on Income from Capital," *Journal of Political Economy*, December 1976, *84*, 1261–83.

Smith, James, *Female Labor Supply: Theory and Estimation*, Princeton: Princeton University Press, 1980.

Solow, Robert, "Capital, Labor and Income in Manufacturing," in *The Behavior of Income Shares: Selected Theoretical and Empirical Issues*, NBER Studies in Income and Wealth, Vol. 27, Princeton: Princeton University Press, 1964, 101–28.

Stuart, Charles, "Swedish Tax Rates, Labor Supply, and Tax Revenues," *Journal of Political Economy*, October 1981, *89*, 1020–38.

______, "Measures of the Welfare Costs of Taxation," mimeo., 1982.

Usher, Dan, "The Private Cost of Public Funds: Variations on Themes by Browning, Atkinson and Stern," mimeo., 1982.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington, 1981.

Internal Revenue Service, *Statistics of Income —1976, Individual Income Tax Returns*, Washington: 1979.

U.S. Department of Commerce, *Survey of Current Business*, Washington, December 1977.

______, *Current Population Reports: Consumer Income*, Washington, 1978.

Money, Credit, and Prices in a Real Business Cycle

By Robert G. King and Charles I. Plosser*

An important recent strain of macroeconomic theory views business cycles as arising from variations in the real opportunities of the private economy, which may include shifts in government purchases or tax rates as well as technical and environmental conditions.[1] These models are often viewed as incomplete or wrong because they do not generate the widely emphasized, but not easily explained, correlation between the quantity of money and real activity.

This paper integrates money and banking into real business cycle theory. The result is a class of models that can account for the correlation between money and business cycles in terms that most economists would label reverse causation.[2] The main focus of the analysis is on the banking system, building on the earlier work of James Tobin (1963) and Eugene Fama (1980). In our real business cycle model, monetary services are privately produced intermediate goods whose quantities rise and fall with real economic developments.

In the absence of central bank policy response, the model predicts that movements in external money measures should be uncorrelated with real activity. Some preliminary empirical analysis (using annual data from 1953 to 1978) provides general support for our focus on the banking system since the correlation between monetary measures and real activity is primarily with inside money.

Our proposed explanation of the correlation between money and business fluctuations stands in sharp contrast to traditional theories that stress market failure as the key to understanding the relation and interpret monetary movements as a primary source of impulses to real activity. Given the controversies surrounding the main contending hypotheses concerning money and business cycles—the incomplete information framework of Robert Lucas (1973) and Keynesian sticky wage models as revitalized by Stanley Fischer (1977)—it seems worthwhile to consider alternative hypotheses.[3]

*Department of Economics and Graduate School of Management, respectively, University of Rochester, Rochester, NY 14627. A preliminary version of this paper was presented at the Seminar on Monetary Theory and Monetary Policy, Konstanz, West Germany, June 1981. We have benefited from the comments of Robert Barro, Herschel Grossman, John Long, Bennett McCallum, anonymous referees, and participants of seminars at the universities of Rochester, Chicago, Pennsylvania, Harvard, and Princeton. The National Science Foundation and the Center for Research in Government Policy and Business of the University of Rochester have supported this research. The above individuals and institutions should not be regarded as necessarily endorsing the views expressed in this paper.

[1] Robert Lucas (1980) provides an overview of the general equilibrium approach to business cycles. Recent work by Fynn Kydland and Edward Prescott (1982) and by John Long and Plosser (1983) illustrate how these models can mimic key elements of business cycles, including complex patterns of persistence and comovement in economic time-series.

[2] The idea that monetary quantities are endogenous is an old one, but has received little recent emphasis. We find it useful to categorize earlier stories into two broad classes: (*i*) banking system explanations such as ours; and (*ii*) explanations that stress central bank policy response. For example, James Tobin (1970) provides an analysis of a model with endogenous money that emphasizes central bank policy response. Tobin's deterministic treatment involves the Keynesian idea that money and real activity respond to the same causal influence—aggregate demand. In Fischer Black's (1972) analysis, external money passively responds to all variations in money demand including those arising from fluctuations in real activity.

[3] In our view, there are good reasons for dissatisfaction with existing macroeconomic theories. Keynesian models typically rely on implausible wage or price rigidities, from the textbook reliance on exogenous values to the recent more sophisticated effort of Fischer (1977) that relies on existing nominal contracts. As Robert Barro (1977) points out, a key feature of the Fischer model is that agents select contracts that do not fully exploit potential gains from trade. In addition, Costas Azariadis' (1978) micro-based model of wage-employment contracts implies that perceived monetary disturbances do not alter output.

Recent analyses of monetary nonneutrality that stress expectation errors based on "imperfect information" (Lucas, 1977, provides a summary of this viewpoint)

The organization of the paper is as follows. In Section I we describe a simple model that is capable of generating real business cycles. The model is used to discuss correlations between an internal monetary quantity and real activity. In Section II, with fiat money included in the model, we analyze the relation between monetary quantities, output, and the price level in both an unregulated and regulated banking environment. In Section III we discuss some of the empirical implications of the theory and provide a preliminary analysis of the postwar U.S. experience.

I. The Real Economy

In this section we describe a simple model economy in which business cycles arise as a consequence of the intertemporal optimizing behavior of economic agents. Our model has two productive sectors with one intermediate and one final good. The output of the final goods industry is stochastic and serves as either a consumption good or as an input into future production. The output of the financial industry is an intermediate good called transaction services that is used by firms in the final goods industry and by households. The demand for transaction services arises because these services economize on time and other resources required to accomplish the exchange of goods.

Recent real general equilibrium theories of the business cycle (such as Finn Kydland and Edward Prescott; John Long and Plosser) stress produced inputs and interrelations between sectors as central to understanding the persistence and comovement of macroeconomic time-series. The simple model economy that we study has only one final product and thus does not possess such a rich set of dynamics or sectoral interactions. Nevertheless, the framework embodies our view that the output of the financial-banking industry is an input into production and purchase of final goods. This view is consistent with the general focus on produced inputs and sectoral interactions that is the hallmark of real business cycle models.

A. *Final Goods Industry*

The single final product (y) is produced by a constant returns to scale production process that uses labor (n), capital (k), and transaction services (d) as inputs. The production technology is summarized by

$$(1)\qquad y_{t+1}=f(k_{yt},n_{yt},d_{yt})\phi_t\xi_{t+1},$$

where k_{yt} is the amount of capital, n_{yt} is the amount of labor services, and d_{yt} is the amount of transaction services used in the final goods industry. Capital services are measured in commodity units allocated to production at time t, labor services are hours worked, and transaction services can be viewed as the number of bookkeeping entries made (described more fully below). We also make the standard assumptions of positive and diminishing marginal products to each factor. The production process is subject to two random shocks, ϕ_t and ξ_{t+1}, that are dated by the time of their realization.

Transaction services in (1) are viewed as an intermediate good purchased by final good producers from the financial industry (to be described below). Although not involved directly in the production of output in the same sense as labor and capital, transaction services are part of a cost-reducing activity similar to other organizational and control inputs.

The sequences $\{\phi_t\}$ and $\{\xi_t\}$ are assumed to be strictly positive stationary stochastic processes that are mutually and serially independent with $E(\phi_t)=E(\xi_t)=1$. The roles played by the two shocks are quite different. At this point it is sufficient to note that ϕ_t alters *expected* time $t+1$ output and affects time t input decisions by altering intertemporal opportunities. On the other hand, ξ_{t+1} represents the basic uncertainty of the production process by altering output in an *unexpected* manner. The multiplicative na-

similarly rely on an apparent failure in the market for information. For example. information on monetary statistics is cheap and readily available. King (1981) demonstrates that in Lucas' (1973) model, real output should be uncorrelated with contemporaneously available monetary information. John Boschen and Herschel Grossman (1982) empirically investigate this proposition and find that it is rejected by the data.

ture of the randomness in total production implies a technological neutrality of the shocks with respect to individual factors of production. Alternatively, different stochastic elements could be associated with particular factors.

Production is assumed to be under supervision of identical competitive firms. Firms operate by selling claims against the future output and using the proceeds to purchase factors of production. Labor, capital, and transaction services are rented at rental prices w_t, q_t, and ρ_t, respectively. Each firm is assumed to sell one unit of claim for each unit of expected output as determined by $f(k_{yt}, n_{yt}, d_{yt})$, which amounts to defining a "share" in the firm. If the market price of claims is v_t, the firm faces a static maximization problem involving the choice of inputs that maximizes profits, $v_t f(k_{yt}, n_{yt}, d_{yt}) - w_t n_{yt} - q_t k_{yt} - \rho_t d_{yt}$. The assumption of constant returns to scale implies that the firm has a supply of claims that is horizontal at the price v_t^*, corresponding to minimum unit cost at prices q_t, w_t, and ρ_t.

B. *Financial Industry*

The financial industry provides accounting services that facilitate the exchange of goods by reducing the amount of time and other resources that otherwise would be devoted to market transactions. The production of this intermediate good, which we call transaction services, is summarized by the production function (2) in which n_{dt} and k_{dt} are the amounts of labor and capital allocated to the financial sector:

$$d_t = h(n_{dt}, k_{dt})\lambda_t. \tag{2}$$

This instantaneous production structure embodies the hypothesis that production of transaction services requires less time than production of the consumption-capital good. Technological innovation in this industry is captured by λ_t, which is assumed to be a strictly positive stochastic process with a mean of one. Finally, we assume (2) represents a constant returns to scale structure so that, at given factor prices w_t and q_t, the financial industry has a supply curve that is horizontal at a particular rental price, ρ_t^*.

Although at this stage of our analysis we focus on the flow of transaction services, the transaction (banking) industry typically (but not necessarily) provides these services in conjunction with portfolio management or intermediary services. It is convenient to imagine, therefore, that the financial industry holds claims (shares) on the probability distribution of output and issues other claims (deposits). In the process of market exchange, the claims that individuals and firms hold on the bank's portfolio (deposits) are altered through simple bookkeeping entries. Banks pass on to depositors the return to the portfolio of assets less a fee for transaction services.

The structure of the financial industry implies that the direct cost of bookkeeping services, ρ_t, does not depend on the character or composition of the bank's portfolio. As discussed by Fama (1980), it follows that there is no reason to expect homogeneous deposits in an unregulated financial industry. More generally, this conclusion holds so long as the respective portfolio costs and transaction services are borne by portfolio holders and transaction users.

C. *Households*

The individual households in the model are consumers, suppliers of labor services and capital goods, purchasers of transaction services, and ultimate wealth holders. The representative individual is assumed to be infinite lived and possess the intertemporal utility function,

$$U_t \equiv \sum_{j=0}^{\infty} \beta^j u(x_{t+j}, \bar{n} - n_{t+j}), \tag{3}$$

where β is a fixed utility discount factor and $u(\cdot)$ is a single period utility function that depends on consumption (x_{t+j}) and leisure $(\bar{n} - n_{t+j})$ with $\bar{n}$ indicating the total hours available in each period. The utility maximand is the expected utility measure $E_t U_t$, where E_t denotes the conditional (rational) expectation based on all information available at time t.

The representative agent arrives at date t with total wealth equal to the sum of current realized output (y_t) and the depreciated value of the previous period's capital stock ($k_{t-1} - \delta k_{t-1}$). The agent's current decisions involve the selection of the levels of consumption (x_t) and of total effort (n_t) as well as allocation of effort to market and nonmarket activities. These decisions imply a level of saving that then must be efficiently allocated, along with current wealth, to purchases of investment goods (i_t) and financial assets (for example, real bonds, shares, etc.).

Households are assumed to combine time and transactions services to accomplish purchases of consumption and investment goods. In particular, the time required for this nonmarket activity is

$$(4) \qquad n_{\tau t} = \tau(d_{ht}/(x_t + i_t))(x_t + i_t),$$

where $\tau' < 0$, $\tau'' < 0$. Our individual selects an amount of transactions services d_{ht} so as to minimize the total transactions cost, $w_t n_{\tau t} + \rho_t d_{ht}$. So long as hours are freely variable, w_t is the opportunity cost of effort, and this minimization problem can be treated separately from the household's general allocations. (However, efficiently selected transactions patterns will have wealth and substitution effects on desirable household allocations.)

Minimizing the total cost of transactions activities implies a derived demand for purchases of transaction services of the form $d^*_{ht} = g(\rho_t/w_t)(x_t + i_t)$, where $g' = (\tau'')^{-1} < 0$. Similarly, hours allocated to transactions activities are proportional to expenditures, taking the form $n^*_{\tau t} = \tau(g(\rho_t/w_t))(x_t + i_t)$.

The presence of transaction costs for the purchase of consumption and investment goods implies that the total cost of a unit of consumption or investment goods in terms of a unit of output is greater than unity (i.e., $1 + [w_t\tau(g(\rho_t/w_t)) + \rho_t g(\rho_t/w_t)]$). Selection of an optimal pattern of consumption (x_t), total effort ($n_t = n_{yt} + n_{dt} + n_{\tau t}$), and portfolio allocations involves the usual sort of intertemporal efficiency conditions with the exception of this modification. Fischer (1982) provides an interpretation of the altered efficiency conditions in a similar context.

D. *Equilibrium Prices and Quantities*

Analysis of dynamic, stochastic general equilibrium models is a difficult task. One strategy for characterizing equilibrium prices and quantities is to study the planning problem for a representative agent (see Lucas, 1978, or Long and Plosser). This procedure is valid so long as the competitive equilibrium is Pareto optimal. The planning problem can also be used to generate specific equilibria if explicit functional forms for preferences and technologies are assumed.

We do not pursue this strategy in detail as our objective is more modest. Instead, we make a number of simplifying assumptions regarding the general framework proposed above that allow us to highlight the conditions necessary to obtain certain business cycle comovements in general equilibrium.

The state of the economy at date t is summarized by the values of four variables; y_t, $(1-\delta)k_{t-1}$, ϕ_t, and λ_t. The first is a measure of national income, the second is the current stock of depreciated capital, ϕ_t is a technical factor affecting current opportunities to transfer resources intertemporally, and λ_t is a technical factor influencing the production of transaction services. The agent's vector of decisions variables is (n_{yt}, n_{dt}, $n_{\tau t}$, d_{ht}, d_{yt}, k_{yt}, k_{dt}).

In order to simplify the problem, we make three assumptions that are sufficient to reduce the state vector to two elements and the decision vector to two elements while preserving the essential features of the model. First, we assume a depreciation rate of 100 percent, eliminating $(1-\delta)k_{t-1}$ as a state variable. Second, we assume that transaction services are produced deterministically ($\lambda_t = 1$, for all t) and depend only on labor input ($d_t = h_0 n_{dt}$). Deterministic production of transaction services eliminates λ_t as a state variable and the simplified production technology implies that the competitive price is $\rho^*_t = w_t h_0$. This implies that households (and firms below) use time and purchased transaction services in fixed proportions.

The third assumption is to restrict the final goods production function to employ financial services in a manner that is symmetric to households. This means that firms (like households) purchase transaction services, d_{yt}, and allocate labor services to transaction activities in fixed proportions where the scale variable corresponds to total payments to factors of production and thus is closely related to next period's output (for households the scale variable is $x_t + i_t$). The second and third assumptions eliminate n_{dt}, $n_{\tau t}$, d_{ht}, d_{yt}, and k_{dt} from the vector of decision variables.

There is a discounted dynamic programming problem whose solution corresponds to the competitive equilibrium of this simplified model economy. The decision rules for the problem are stationary functions of the state variables y_t and ϕ_t. Rather than solve this problem for an explicit specification of preferences and technologies, the essential features of the interactions between the final goods industry and the financial industry can be analyzed by employing the following restrictions on the decision rules; $0 < \partial k_{yt}/\partial y_t < 1$, $\partial k_{yt}/\partial \phi_t \cong 0$, $\partial n_{yt}/\partial y_t > 0$, and $\partial n_{yt}/\partial \phi_t > 0$.

These restrictions follow from assumptions about preferences and production opportunities. For example, an increase in the amount of the initial stock, y_t, involves additional wealth so that the consumption of final product and leisure are expected to rise. Agents, however, choose to spread some portion of this wealth increment over time and do so by increasing the amount of commodity allocated to capital services so that $0 < \partial k_{yt}/\partial y_t < 1$. The other conditions on the decision rules require stronger restrictions on preferences and production possibilities. For example, an increase in y_t raises the marginal product of labor if capital and labor are complements in production. If the wealth effect on labor supplied, which arises from the increased output of final goods next period, is outweighed by the increase in the real wage (marginal product of labor) then hours worked rises.[4]

Analogously, an increase in ϕ_t involves both wealth and substitution effects. Given current inputs, future production is higher and the current returns to additional units of factors of production are higher. These offsetting effects are analogous to the income and substitution effects of a real interest rate change. Essentially, the small impact of a shift in ϕ_t on the amount of output allocated to capital accumulation ($\partial k_{yt}/\partial \phi_t \cong 0$) reflects the idea that the income and substitution effects on consumption are roughly offsetting. On the other hand, the substitution effect of such shifts on labor supply is presumed to dominate so that $\partial n_{yt}/\partial \phi_t > 0$, which generates procyclical work effort.

Once quantity behavior is determined, equilibrium factor prices, interest rates, and share prices are straightforward to construct. In particular, competitive prices correspond to marginal rates of substitution at optimal planned quantities. For example, there is a riskless commodity interest rate r_t that we discuss below. We also can construct the expected return to shares, $E_t(r_{yt}) = E_t[\phi_t \xi_{t+1} - v_t)/v_t] = (\phi_t - v_t)/v_t$. In our setup, this expected return exceeds the riskless rate, $E_t(r_{yt}) > r_t$, since the holders of these shares must be compensated for bearing production risk.

E. *Inside Money, Credit, and the Real Business Cycle*

In our real business cycle model, a positive correlation (comovement) of real production, credit, and transaction services arises from the general equilibrium of production and consumption decisions by firms and households. The timing patterns among these variables, however, depends on the source of the variation in real output.

Unexpected output events (ξ_t) operate by altering the initial conditions pertinent for economic agents' plans for consumption, investment, and hours of work. As discussed above, an unexpected wealth increment ($\xi_t > 1$) leads to higher net investment than would otherwise have been the case. Furthermore, hours worked also rises so that real output increases and exhibits positive serial correlation. During the course of such an economic expansion, the volume of credit (shares) is also high as firms finance relatively large

[4] This result also requires that the amount of time allocated to transaction activities by firms and household is small relative to total time allocated to market activity or production.

amounts of goods in process. This positive correlation between the total volume of credit and real activity is potentially an important prediction of our framework, especially since evidence presented by Benjamin Friedman (1981) suggests that there is a tighter relation between total credit and output than between the individual components of credit and real activity.

The movements in final goods production induces a higher volume of transaction services demanded by firms and households. Thus, our model generates the positive comovement of output with measures of bank clearings, long noted by empirical researchers in the business cycle area (for example, Wesley Mitchell, 1930, pp. 116–51). Finally, real rates of return move in a countercyclical direction as agents' opportunities to spread wealth over time are subject to diminishing returns (i.e., total time is in fixed supply).

The predictions of our model focus on the flow of transaction services. It is important to provide a link between these flows and the stock of deposits that has been the more traditional focus of monetary analysis. It is convenient to assume that the stock of deposits is proportional to the flow of transaction services and can be represented by γd_t.[5] Under this assumption, our model implies that the volume of inside money (deposits) is positively correlated with output with a rough coincidence in timing. More generally, this may reflect the role of deposits as a store of wealth or a temporary element of the credit process.

At least some cycle episodes, however, are commonly viewed as involving a different timing pattern. For example, traditional business cycle analysts (Arthur Burns and Mitchell, 1946), modern time-series macroeconometricians (Christopher Sims, 1972; 1980),[6] and monetary historians (Milton Friedman and Anna Schwartz, 1963) view monetary variables as "leading" measures of real activity.[7]

One way of generating a different timing pattern is through shifts in the intertemporal opportunities of the economy as a whole. Real events of this type, respresented by ϕ_t, alter agents' allocations of leisure and consumption between the present and the future for a given level of national wealth. A higher than average shock ($\phi_t > 1$), under the assumptions outlined above, expands hours worked with little accompanying change in consumption or capital. The fact that financial services are an intermediate product—which can be produced more rapidly than the final product—leads to an expansion of the quantity of such services and of bank deposits. Consequently, movements in hours worked, interest rates, and security prices, deposits-financial services, and trade credit all occur prior to the expansion of output.[8] The subsequent increment to time $t+1$ wealth (stemming from the joint impact of the exogenous shift, ϕ_t, and agents' responses to that shift) works much like the above discussion of unexpected output events. Typically, we suspect the initial phases of business fluctuations incorporate a combination of both types of shocks (i.e., shifts in current and expected future production possibilities).

II. Currency, Deposits, and Prices

In order to investigate the relation between nominal aggregates and the real business cycle it is necessary to augment the

[5] It is sufficient for our purposes that deposits be related to transaction services by any monotonic increasing function. Although this assumption is a conventional assumption with physical capital, it is nevertheless an *ad hoc* element that is troubling. For example, transaction services do not, in principle, require any specific asset position, as is clear from checking accounts that have overdraft privileges or carry a zero balance at the end of the day. In addition, there are important secular and cyclical variations in the volume of debits relative to the stock of deposits.

[6] Sims (1980) discusses reverse causation of money and output working through central bank operating policies. The present setup is a first step toward the type of small-scale general equilibrium model that is necessary to evaluate the reverse causation argument.

[7] We deliberately employ the idea of a "leading variable" in a loose manner so as to capture the common elements of these alternative discussions.

[8] It is commonly stressed that asset prices and returns incorporate information about predictable components of future output (i.e., ϕ_t). In general equilibrium models such as ours, however, such information is also incorporated into all quantity decisions such as effort, consumption, and investment.

hypothetical economy developed above. In this section, a non-interest-bearing government-supplied fiat currency (dollars) is introduced and the factors affecting its value are analyzed.

In order for currency to be a well-defined economic good, and thus to have a determinate price in terms of a unit of output $(1/P)$, there must be a demand function for currency that reflects the economic value assigned to the services of currency by economic agents. For simplicity, we assume that households are the principal demanders of currency. To generate a stable demand for currency, real currency is viewed as a substitute—but not a perfect one—for transactions services purchased from the financial sector. In particular, currency yields a real service flow in that there are some transactions (either of magnitude or character) that are more efficiently carried out using currency than the accounting system of exchange. We revise the household's time spent in transactions activities, equation (4), to reflect these expanded opportunities:

$$n_{\tau t} = \tau(d_{ht}/y_t, c_t/y_t)y_t, \tag{5}$$

where y_t is the total market transactions of our household, c_t is the stock of currency purchasing power, and d_{ht} is the flow of financial services purchased from the financial industry.[9]

Thus, a household minimizing its cost of transaction activities will select amounts of $n_{\tau t}$, d_{ht}, and c_t in a manner that is analogous to our earlier discussion. The demands for each input will be a function of the rental price of real currency, $R_t/(1+R_t)$, the effective cost of financial services $\bar{\rho}_t$ and the opportunity cost of time, w_t.

In forming these rental prices, two important assumptions are made. First, there is a market for one-period nominal bonds that bear interest rate R_t. This nominal rate is the sum of the real component, r_t, and an expected rate of inflation, $E(\pi_t)$. Second, if banks are required to hold non-interest-bearing reserves, the returns earned by depositors may not match market rates. Given the proportional link between financial services (d_t) and deposits (γd_t), the effective cost of a unit of deposit services, $\bar{\rho}$, is influenced by this reserve regulation. In the absence of regulations $\bar{\rho}_t = \rho_t$, where ρ_t is the rental price of deposit services in the competitive environment of Section I.

Transactions cost minimization by households implies real demands for currency and financial services of the following forms,

$$c_t = l(\underset{-}{R_t/(1+R_t)}, \underset{+}{\bar{\rho}_t}, \underset{+}{w_t})y_t, \tag{6a}$$

$$d_{ht} = \delta(\underset{+}{R_t/(1+R_t)}, \underset{-}{\bar{\rho}_t}, \underset{-}{w_t})y_t. \tag{6b}$$

The signs below the arguments denote the signs of the partial derivative (for example, $\partial c_t/\partial \bar{\rho}_t > 0$). These signs are insured if $\tau(\cdot)$ is such that currency purchasing power and financial services are substitutes.

The structure of the markets for currency and financial services is analogous to Fama (1980). As he points out, determinacy of the price level is insured if the government fixes the nominal quantity of currency—direct or indirect regulation of financial sector quantities and/or characteristics is not necessary. Nevertheless, regulations can be important for two reasons. First, regulations produce a differentiated class of suppliers of financial services (banks) whose deposits are sometimes described as inside money. Second, regulations can influence the price level by altering the effective rental price of financial services.

The analysis below focuses on the implications of alternative banking structures for the behavior of currency, deposits, and prices. For clarity, the bulk of the discussion is conducted under the assumption that the treasury-central bank maintains a policy of controlling the issue of nominal currency so that the stock of currency $(C_t = P_t c_t)$ is an exogenous random variable. This assumption means that the behavior of the price level can be analyzed by investigating equilibrium in the currency market. Other models of central bank behavior are discussed in Part B below. In all of our discussions, however, the price level is best viewed as being set in the

[9] We assume that $\tau_1 < 0$, $\tau_2 < 0$, $\tau_{11} < 0$, $\tau_{22} < 0$, and $\tau_{12} < 0$.

market for that nominal asset whose quantity the central bank seeks to control.

A. *Money and Prices — Unregulated Banking*

In an unregulated banking environment we assume that the deposit industry would hold virtually no currency. Consequently, the determination of the price level involves the requirement that the real supply of currency (C_t/P_t) be equal to the real demand for currency given by (6a) above. The equilibrium price level is then

$$P_t = C_t / l(\cdot), \tag{7}$$

where $l(\cdot)$ is the demand function for real currency. Using the arguments of $l(\cdot)$ we can rewrite this condition as

$$P_t = P(\underset{+}{C_t}, \underset{-}{y_t}, \underset{+}{R_t}, \underset{-}{\bar{\rho}_t}, \underset{-}{w_t}). \tag{8}$$

The signs of the respective derivatives in (8) are straightforward and warrant little explanation.

An important feature of (8) is the absence of nominal demand deposits. Thus, as stressed by Fama (1980), there is no need for government control of banking or the supply of deposits to insure a determinate price level. Banks, in a competitive, unregulated environment, simply pass portfolio returns on to their depositors less a fee charged for the provision of transactions services, so that $\bar{\rho}_t = \rho_t$. The only way in which developments in the banking sector are relevant to price level determination is through variations in the cost of financial services (ρ_t).

This view of price level determination implies that once and for all changes in the quantity of currency are completely neutral. The volume of transaction services (d_t) and deposits (γd_t) are determined solely by variations in the real economy as discussed in Section I. Nevertheless, the *nominal* quantity of deposits $(P_t \gamma d_t)$ is likely to be positively correlated with real activity if currency is determined exogenously and prices are not excessively countercyclical (see Part C below).

On the other hand, sustained increases in the growth of currency may have real effects. The resulting increased inflation leads to a rise in the nominal interest rate, R, which implies a fall in the demand for real currency and a rise in real transaction services and time allocated to transaction activities. Since an increase in real transaction services involves the use of real resources, the economy is made worse off by sustained inflation. We assume, however, that this increase in the size of the financial sector has no important implications for the real general equilibrium.[10] It is not obvious that this is a good assumption from an empirical point of view. Nevertheless, it does serve to bring into sharp focus the distinction between inside and outside money, particularly with respect to the neutrality and super-neutrality of government currency issue.

B. *Money and Prices — Regulated Banking*

In an unregulated environment the price level is determined in the currency market and deposits play no essential role. There are, however, a number of regulations that serve to distinguish banks from other financial intermediaries and thereby inside money from credit. Here we discuss the extent to which these regulations alter the nature of price level determination. As it turns out, the impact of regulations depends on (*i*) the interaction of banking regulation with the external money supply policy of the central bank-treasury, and (*ii*) the extent to which government mandates can be offset by countervailing private substitutions.

1. *Portfolio Regulations and Reserve Requirements.* It is useful to start by discussing a set of regulations that do not have any important consequences for the price level.

[10] In other words, we assume that our model is approximately "super-neutral" in the language of monetary growth theory. It is worthwhile pointing out that this literature does not provide a clearcut guide to the nature of departures from super-neutrality. For example, Tobin (1965), has argued that an increase in inflation will lower real rates of return and raise capital formation, by lowering the real value of money and, consequently, raising saving. By contrast, Alan Stockman (1981) argues that inflation acts as a tax on the saving process (in which money is an input) and, hence, depresses capital formation.

Suppose that the government specifies the "risk composition" of the underlying assets against which deposits are claims. As long as agents can offset this restriction by rebalancing the contents of their portfolios (i.e., the distribution of total wealth between the banking sector and other portfolio managers), then this regulation will have no impact on any real variables or the price level. However, such restrictions may serve to distinguish inside money for other forms of credit.

On the other hand, restrictions specifying that banks must hold some fraction, say θ, of their nominal asset portfolio in the form of non-interest-bearing reserves issued by the central bank may have important effects. For example, the central bank could specify that reserve accounts are deposits of securities with nominal interest accruing to the central bank. This mechanism is one way of imposing a deposit tax with the consequence that the cost of deposit services would be $\bar{\rho}_t > \rho_t$. Such a deposit tax results in a reduction in the size of the banking sector and an increase in the real demand for currency. The impact of this reserve requirement on price-level determination depends on the central bank policy. For example, if the treasury-central bank makes currency in the hands of the public an exogenous quantity, unresponsive to developments in the banking sector, then the price level continues to be determined by the requirement that the real stock of currency outstanding (C_t/P_t) be equal to the real demand. In these circumstances, the behavior of deposits and deposit services would be similar to that in an unregulated banking system.

2. *Alternative Central Bank Policies.* The currency market determines the price level if the central bank is assumed to make currency an exogenously controlled quantity. There are, however, other control methods available to the central bank. For example, if the central bank combines a reserve requirement with a policy of controlling the sum of currency and nominal bank reserves (high-powered money), then the price level can be viewed as being determined in the market for high-powered money.

Let $B_t = \theta(P_t \gamma d_t)$ be the nominal stock of bank reserves and $H_t = B_t + C_t$ be the exogenous total of bank reserves and currency. Under this regime, the price level may be viewed as arising from the requirement that the total private demand for fiat money equals the supply. That is, $H_t = P_t\{c_t + \theta\gamma d_t\} = P_t\{c_t + B_t/P_t\}$. The equilibrium price level can be expressed as

$$(9) \qquad P_t = H_t/\big(l(\cdot)+(B_t/P_t)\big),$$

or using the arguments of $l(\cdot)$,

$$(10) \qquad P_t = P(\underset{+}{H_t}, \underset{-}{y_t}, \underset{+}{R_t}, \underset{-}{\bar{\rho}_t}, \underset{-}{w_t}, \underset{-}{(B_t/P_t)}).$$

Once again the signs of the partial derivatives are straightforward. Note, in particular, that an increase in the demand for real reserves (B_t/P_t) holding high-powered money fixed necessitates a fall in the price level.

A central bank policy of controlling high-powered money, therefore, implies that the equilibrium price level is determined in the market for this exogenously controlled nominal quantity. Consequently, real activity (including real deposit services and real deposits) is neutral with respect to changes in high-powered money and, as in the case of currency, we assume that high-powered money is approximately super-neutral under this regime.

As discussed in Fama (1980, pp. 52–53), there are other central bank policies that could be used to make the price level determinate. In particular, the central bank could choose to make nominal bank reserves an exogenous quantity and supply currency on demand. In this case (which some argue are the current policies of the Federal Reserve), the price level can be viewed as being determined in the market for reserves. The equilibrium price level would be determined by the exogenous supply of nominal reserves (B_t) and the total real demand for deposit services.[11]

[11] We have not yet analyzed price-level determination when the central bank attempts to control the interest rate. However, this may be important to an appropriate empirical investigation of some time periods.

C. *The Price Level and the Real Business Cycle*

Price-level movements in response to the two shocks (ϕ_t and ξ_t) involve two important factors. First, there is the impact of movements in real output on the demand for outside money. Second, there is the impact of nominal interest rates on the demand for outside money. Since variation in the price level also depends on central bank policy, we focus on the case of a regulated banking system with the central bank assumed to make the quantity of high-powered money exogenous.

It is convenient to summarize household and bank behavior in the following demand function for outside money:[12]

$$h_t^d = p_t + \lambda y_t - \psi R_t, \qquad \lambda > 0, \psi > 0,$$

where h_t is the logarithm of high-powered money, y_t is the logarithm of real output, p_t is the logarithm of the price level, and R_t is the nominal interest rate. Using the fact that $R_t = r_t + (E_t p_{t+1} - p_t)$ and the monetary equilibrium condition that $h_t = h_t^d$, it follows that a rational expectations solution for the price level along the lines of Thomas Sargent and Neil Wallace (1975) can be written as

$$p_t = (1+\psi)^{-1}\left\{\sum_{j=0}^{\infty} (\psi/(1+\psi))^j \cdot E_t\left[h_{t+j} + \psi r_{t+j} - \lambda y_{t+j}\right]\right\}.$$

Unexpected wealth increments ($\xi_t > 1$) lead to a business cycle where output is high and the real rate of return is low. Consequently, a wealth increment leads to lower prices due to both lower real returns and higher income.

In Section I we describe how a better than average opportunity to transfer resources intertemporally ($\phi_t > 1$) leads to an increase in r_t. In addition, the increase in wealth that is brought about by such a shift leads to lower future returns and higher future outputs. Thus, the overall impact on the price level is ambiguous, involving the positive influence of the higher current real return and the negative influence of the lower expected future returns and higher expected future outputs.

The above two examples suggest that the model produces a price level that is likely to be countercyclical. For some macroeconomists, the procyclical character of the general price level is such a well established empirical regularity that this feature alone is sufficient to reject real business cycle theory (for example, Lucas, 1977, p. 20).[13] If it is indeed necessary to generate procyclical price movements, then there appear to be two principle channels. First, an alternative structure that involves a more permanent, capital-augmenting form of technological change could heighten the real return effects discussed above. Combining this structure with a sufficiently interest-sensitive demand for money could lead to procyclical prices. Second, policy response to real activity also could generate procyclical price movements. For example, a positive response of outside money creation to output could lead to a positive correlation between prices and output.

III. Empirical Analysis

The preceding sections describe a simple model economy with business cycles that are completely real in origin. Nevertheless, correlations between real activity and monetary measures arise from the operation of the banking system and central bank policy responses. Here we discuss some of the predictions that our model makes concerning the joint time-series behavior of output, monetary aggregates, rates of return, and the price level. In addition, we discuss U.S. business cycle experience during the post-World

[12] For simplicity, we ignore movements in the cost of deposit services and real wage as important factors affecting the price level.

[13] Recent work using post-World War II data (for example, Robert Hodrick and Prescott, 1980, and Fama, 1982) suggests that the positive correlation between output and price level movements may be not as robust as sometimes thought.

War II period, providing some preliminary empirical evidence that bears on the potential relevance of our theoretical stories.

Before proceeding, it is useful to consider briefly general strategies for investigating the empirical importance of real business cycle theories, and to discuss how the present analysis of money and the price level could be related to such investigations.

One empirical strategy is to isolate a group of observable real disturbances that provide an explanation of much of a particular nation's business cycle experience, in the sense of delivering a good fit. Candidates for such real shocks include government purchase, tax, and regulatory actions; changes in technological and environment conditions, and movements in relative prices that are determined in a world market. The goal is to provide a direct substitute for the high explanatory power of monetary variables found in other business cycles studies (for example, Friedman and Schwartz, and Barro, 1981a). The natural extension would be to study the explanatory power of such real factors for monetary variables and the price level. In our framework, many of these real variables would be restricted to influence monetary quantities (particularly, inside money) through their influence on output and a small set of relative prices. In this sense, aspects of the present type of monetary theory do provide meaningful restrictions on the data.

Another approach is to treat the fundamental real shocks as unobservable and to focus on the interactions between sectors that arise during business cycles; a strategy that is the empirical analogue of the theoretical analysis of Long and Plosser. Since a particular real business cycle theory restricts own- and cross-serial correlation properties of industry output and relative prices, this route can provide valuable information about the regular aspects of business cycles even though the sources of shocks are not identified. Again, the principal testable restrictions of theorizing along these lines would arise from the restricted fashion that variations in production in other sectors were allowed to influence developments in the monetary sector.

Unfortunately, analysis of monetary phenomena using either of these strategies is not feasible given the state of real business cycle models. Consequently, the present empirical investigation is limited to providing some admittedly crude correlations among the variables suggested by the theory.

A. *Summary Statistics*

Summary measures of the series to be discussed below are presented in Table 1. The data are annual (generally yearly averages) for the period 1953–78. We focus on the 1953–78 interval primarily to avoid the period when the Federal Reserve maintained a policy of pegging the yields on U.S. government securities. The implications of such a policy may be very different from those described in the previous section where the central bank controls some nominal quantity.

The most noticeable feature in Table 1 is the different behavior of nominal and real variables. Typically, the growth rates of real variables display much less serial correlation than the growth rates of nominal variables. For example, the growth of real demand deposits is much less autocorrelated than the growth rate of nominal demand deposits. Indeed, as previously noted by Charles Nelson and Plosser (1982), as well as other authors, many real variables are close to random walks in logarithmic form. The most noticeable exceptions to this random walk behavior are real currency (C_t/P_t) and real service charges (ρ_t), both of which display significant positive and persistent serial dependences in growth rates.

B. *Real Factors and Aggregate Output*

This paper is not the appropriate place for an empirical investigation of the role of real factors as impulses to business fluctuations. Barro (1981b), however, provides some results that are pertinent. Specifically, he finds that temporary increases in government purchases have a significant expansionary impact on real output. These results are suggestive and one could investigate the impact of other governmental tax and expendi-

TABLE 1—SUMMARY STATISTICS, ANNUAL DATA: 1953–78

Series	Mean	Standard Deviation	ρ_1	ρ_2	ρ_3	ρ_4
A. Real Variables						
Growth Rate of Real						
GNP (y_t)	.0327	.0249	−.01	−.24	−.12	.29
Wages (w_t)	.0177	.0324	.59	−.00	−.02	.14
Deposits (γd_t)	−.0002	.0226	.36	−.22	−.19	.20
Currency (C_t/P_t)	.0101	.0209	.65	.39	.26	.25
High-Powered Money (H_t/P_t)	.0027	.0253	.40	.33	.08	.19
Reserves (B_t/P_t)	−.0110	.0449	.32	−.01	−.05	−.02
Service Charges (ρ_t)	−.0252	.0601	.81	.69	.66	.56
B. Nominal Variables						
Growth Rate of						
Price Level (P_t)	.0371	.0233	.84	.64	.66	.84
Deposits ($P_t\gamma d_t$)	.0373	.0211	.58	.35	.43	.58
Currency (C_t)	.0481	.0329	.93	.88	.85	.82
High-Powered Money (H_t)	.0398	.0338	.71	.76	.59	.68
Reserves (B_t)	.0260	.0455	.37	.03	.09	.32
Change in the Short-Term Interest Rate (R_t)	.2177	1.4710	.03	−.71	−.29	.68

Note: ρ_i is the sample autocorrelation coefficient at lag i, for $i = 1,\ldots,4$. The large sample standard error is .20.

Sources: Real *GNP* and the *GNP* deflator are taken from *The National Income and Product Accounts of the United States, 1929–1941* and various issues of the *Survey of Current Business*. Currency in the hands of the public, demand deposits, and bank reserves are from *Business Statistics, 1979*. High-powered money is the sum of currency in the hands of the public and bank reserves. The interest rate is the 4- to 6-month prime commercial paper rate taken from *Banking and Monetary Statistics 1941–1970* and various issues of the *Annual Statistical Digest*. The real wage is average hourly earnings divided by the producer price index, from *Business Statistics, 1979*. Finally, the service charge variable is the ratio of total service charges on demand deposits accrued by Federal Reserve member banks to total check clearings by the Federal Reserve. Both series are taken from *Banking and Monetary Statistics, 1941–1970*, and various issues of the *Annual Statistical Digest*.

ture measures on real activity. More recently, David Lilien (1982) documents the importance of a measure of the dispersion of sectoral shifts in understanding the movements in aggregate unemployment during the postwar period. James Hamilton (1983) presents evidence on the relation between oil price changes and postwar recessions.

Additional evidence on the importance of real disturbances in output fluctuations is offered in Nelson and Plosser. Using an unobserved components model of output and the observed autocovariance structure of real *GNP*, Nelson and Plosser infer that real (nonmonetary) disturbances are the primary source of variance in real activity. This result is based on the commonly held view that monetary disturbances should have no permanent effects on real output, and thus disturbances that are of a permanent nature must be associated with real rather than monetary sources.

C. *Money-Output Correlations*

The theoretical model stresses that real internal monetary balances should be positively correlated with real activity since transaction services are a produced input. Further, the model predicts that autonomous external nominal money creation/destruction is neutral with respect to output growth. These two ideas suggest the value of analyzing money-output correlations in two forms: real vs. nominal balances and internal vs. external monetary measures.

Table 2 presents information on the contemporaneous relations between output growth and growth rates of alternative monetary measures. Equation (i) shows the strong positive contemporaneous correlation that exists between real demand deposits and economic activity. This strong contemporaneous correlation is shared by real external balances measured as currency or as high-

TABLE 2—CONTEMPORANEOUS MONEY-OUTPUT REGRESSIONS

$$\Delta \ln y_t = \alpha_0 + \alpha_1 \Delta \ln M_t + \varepsilon_t$$

		Independent Variables (M_t)								
		Real Monetary Measures			Nominal Monetary Measures					
Equation	$\hat{\alpha}_0$	γd_t	H_t/P_t	C_t/P_t	$P_t \gamma d_t$	H_t	C_t	R^2	$SE(\hat{\varepsilon})$	ρ_1
(i)	.033[b] (.004)	.740[b] (.167)						.450	.0188	−.08
(ii)	.031[b] (.004)		.510[b] (.103)					.337	.0206	−.18
(iii)	.025[b] (.005)			.664[b] (.202)				.311	.0211	−.01
(iv)	.015 (.009)				.465[b] (.222)			.155	.0233	.10
(v)	.026[b] (.007)					.171 (.146)		.054	.0247	.00
(vi)	.027[b] (.009)						.111 (.153)	.022	.0251	.01
(vii)	.025[b] (.006)	.742[b] (.161)				.176[a] (.108)		.507	.0182	−.11
(viii)	.023[b] (.006)	.784[b] (.162)					.194[a] (.111)	.514	.0181	−.08
(ix)	.017[a] (.010)				.558[a] (.326)	−.080 (.203)		.161	.0238	.10
(x)	.015 (.010)				.661[b] (.307)		.181 (.197)	.185	.0234	.07

Notes: See Table 1; $\Delta \ln(\cdot)$ indicates the change in the log of the variable; R^2 is the coefficient of determination; $SE(\hat{\varepsilon})$ is the standard error of the regression; ρ_1 is the estimated first-order autocorrelation coefficient of the residuals, which has a large sample standard error of .20. Standard errors of the coefficients are shown in parentheses.

[a] Indicates significance at the 10 percent level.

[b] Indicates significance at the 5 percent level.

powered money (equations (ii) and (iii)). In nominal balance form, equations (iv), (v), and (vi) show demand deposits are more strongly correlated with real activity than either of the nominal external money measures. Finally, (vii) and (viii) indicate that nominal high-powered money and currency growth have a weak positive partial correlation with output given real demand deposits.

From the standpoint of our theoretical discussion, the key aspects of these correlations are as follows. First, the fact that much of the correlation with real activity is with internal monetary measures is consistent with our general view of the relation between money and real activity. Second, the fact that currency or high-powered money may be positively correlated with real activity is at odds with our model so long as the monetary authority makes such nominal monetary measures evolve in an autonomous manner.

Table 3 reports some additional regression results that incorporate lags of the alternative monetary measures. Equation (i) shows the results of adding two years of lagged real deposits to the output regression. The F-statistic pertinent for evaluating the marginal contribution of these lags is 2.48, which is well below the 95 percent critical value of 3.49, so that there is no strong evidence that these lags are important. Equations (ii) and (iii) show analogous results for nominal money growth measures.

Equations (iv) and (v) investigate the extent to which nominal money growth is correlated with real activity after accounting for real deposit growth. The contemporaneous and second lag of high-powered money and currency in the hands of the public are not important explanatory variables (the 95 percent critical value for $F(3,17)$ is 3.20 and the F-statistics for the lags of high-powered

TABLE 3—MONEY GROWTH AND OUTPUT GROWTH REGRESSIONS

$$\Delta \ln y_t = \alpha_0 + \sum_{i=0}^{2} \beta_i \Delta \ln \gamma d_{t-i} + \sum_{i=0}^{2} \gamma_i \Delta \ln H_{t-i} + \sum_{i=0}^{2} \delta_i \Delta \ln C_{t-i} + \varepsilon_t$$

		Independent Variables											
		Real Deposits (γd_t)			High-Powered Money (H_t)			Currency (C_t)					
Equation	$\hat{\alpha}_0$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\delta}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$	R^2	$SE(\hat{\varepsilon})$	ρ_1
(i)	.034^b	.644^b	.135	−.352^b							.651	.0152	.01
	(.003)	(.159)	(.166)	(.159)									
(ii)	.035^b				.489^b	−.399^a	−.175				.236	.0225	.05
	(.008)				(.239)	(.210)	(.243)						
(iii)	.032^b							.342	−.577	.262	.068	.0249	.18
	(.010)							(.487)	(.493)	(.360)			
(iv)	.034^b	.607^b	.066	−.296^a	.263	−.229	−.059				.713	.0150	.00
	(.006)	(.160)	(.189)	(.162)	(.182)	(.145)	(.185)						
(v)	.031^b	.644^b	.119	−.333^a				.302	−.313	.017	.674	.0160	.00
	(.006)	(.173)	(.178)	(.171)				(.317)	(.321)	(.238)			
(vi)	.033	.605^b	.091	−.305^a	.233^a	−.233^a					.710	.0142	.02
	(.003)	(.150)	(.157)	(.151)	(.118)	(.118)							
(vii)	.033	.642^b	.114	−.343^b				.297	−.297		.670	.0152	.00
	(.003)	(.158)	(.167)	(.159)				(.280)	(.280)				

Note: See Table 2. Equations (vi) and (vii) are the results of the regressions that constrains $\gamma_0 = -\gamma_1$ and $\delta_0 = -\delta_1$.
[a,b]See Table 2.

money and currency terms are 1.22 and .40, respectively). However, the estimated coefficient on current and lagged high-powered money are opposite in sign and nearly identical in magnitude, so that the change in high-powered money growth appears to be positively correlated with real activity (see equation (vi)).

Overall, our interpretation is that the correlations reported in Tables 1 and 3 indicate that much of the relation between money and real activity is apparently one with inside money, which is comforting given the key role that the banking system plays in our theoretical story.[14] Nevertheless, somewhat weaker correlations between real activity and nominal outside money may exist, suggesting it may be necessary to analyze policy response in greater detail for the 1953–78 period, or to relax our maintained assumption of super-neutrality.

D. *Money-Inflation Correlations*

The theoretical model predicts that variations in external money, real activity, the nominal interest rate, and a measure of the cost of banking services should be important in explaining movements in the price level. Table 4 provides estimates of the price-level equations (8) and (10) of Section III under the assumption that a log-linear functional form is appropriate. Although the nominal interest rate is endogenous and the above discussion indicates that high-powered money and/or currency may be endogenous due to policy response, ordinary least squares methods are employed. Since there is a substantial empirical literature on price-level/money demand equations, our discussion focuses principally on new aspects that are raised by the theoretical discussion above.

First, the theory suggests that a measure of external money, such as currency or high-powered money, is the relevant nominal ag-

[14]Although we confine our empirical analysis to a comparison of inside and outside money correlations with output, broader measures of financial assets (or credit) should behave similarly to inside money. Friedman presents additional empirical support for this view. He finds that broader measures of money (or credit) exhibit a higher correlation with real output than more narrowly constructed measures.

TABLE 4—INFLATION REGRESSIONS

$$\Delta \ln P_t = \alpha_0 + \alpha_1 \Delta \ln M_t + \alpha_2 \Delta \ln y_t + \alpha_3 \Delta R_t + \alpha_4 \Delta \ln w_t + \alpha_5 \Delta \ln \rho_t + \alpha_6 \Delta \ln(B_t/P_t) + \varepsilon_t$$

Equation	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	R^2	$SE(\hat{\varepsilon})$	ρ_1
A. Currency as External Money										
(i)	.027^{b}	.474^{b}	−.276^{b}	−.000	−.224^{b}	−.014		.842	.0103	.32
	(.005)	(.098)	(.115)	(.002)	(.097)	(.056)				
(ii)	.025^{b}	.457^{b}	−.246^{b}	.001	−.201^{b}	−.029	−.091	.862	.0099	.33
	(.005)	(.094)	(.112)	(.002)	(.094)	(.054)	(.055)			
(iii)	.008	1.00	−.397^{b}	.001	−.134	.178^{b}		.519	.0158	.39
	(.005)		(.173)	(.003)	(.145)	(.065)				
(iv)	.007	1.00	−.381^{b}	.002	−.118	.712^{b}	−.058	.521	.0161	.38
	(.006)		(.177)	(.003)	(.150)	(.067)	(.088)			
B. High-Powered Money as External Money										
(i)	.035^{b}	.334^{b}	−.240	−.002	−.255^{b}	−.070		.753	.0130	.29
	(.005)	(.119)	(.144)	(.002)	(.120)	(.068)				
(ii)	.023^{b}	.498^{b}	−.208^{a}	.001	−.165^{a}	−.035	−.240^{b}	.870	.0096	−.06
	(.005)	(.097)	(.107)	(.002)	(.092)	(.051)	(.058)			
(iii)	.017^{b}	1.00	−.386^{a}	−.004	−.157	.163^{a}		.464	.0202	.07
	(.007)		(.221)	(.003)	(.186)	(.084)				
(iv)	.007	1.00	−.285^{a}	.001	−.057	.129^{b}	−.362^{b}	.735	.0146	−.32
	(.005)		(.161)	(.003)	(.135)	(.061)	(.080)			

Note: See Table 2.
[a,b]See Table 2

gregate for price level determination (Fama, 1982, also advances this hypothesis and provides relevant evidence). Table 4 presents empirical results for both currency and high-powered money.

Second, in the regulated banking environment described in Section II, the relevant cost of deposit services (denoted $\bar{\rho}_t$) involves both the direct cost of providing an accounting system of exchange (denoted ρ_t) and the interest that the bank-depositor must forego due to reserve requirements. The empirical counterpart to the nominal unit cost of deposit services that we have constructed is the ratio of total service charges on demand deposits accrued by Federal Reserve member banks to total check clearings by the Federal Reserve. Deflating this measure by the price level leads to a measure of the real costs of deposit services, entered in Table 4 as ρ_t. However, during some portions of the period under study, banks faced apparently binding constraints on the level of interest payments that could be paid on demand deposits. It is frequently argued that explicit service charges would be reduced as a means of avoiding the interest rate constraint. As a result, we are not completely comfortable with our interpretation of this variable.

Third, our model of transaction costs implies that the real wage is also a pertinent relative price variable for agents in determining the mix of currency and transaction service purchases. As the real wage rises, individuals substitute toward the use of currency and purchased transaction services in market exchange.

Finally, when reserve requirements are present and the central bank is controlling the quantity of high-powered money the theory predicts that the volume of real reserves should negatively influence the price level given the stock of high-powered money. On the other hand, when currency is the controlled external quantity, real reserves should not be relevant.

In Panel A of Table 4, equations (ii) and (iv) report the results of estimating the price level (inverse money demand) equation over the sample period 1953–78, with currency as the measure of external money. The main features of these equations are broadly consistent with other studies: a negative impact of real activity, positive impact of nominal money growth, and minor or negligible impact of the short-term interest rate (4- to 6-month commercial paper rate). Although not included in many other studies, the real wage enters these equations in a manner that is consistent with our theory. If currency is the appropriate measure of external money the theory predicts a zero coefficient on real reserves. In equation (ii) this coefficient is negative but insignificant by the usual criteria. The tendency of our service charge measure to switch sign with the imposition of the unit constraint on currency (iv) is troubling, casting some doubt on the appropriateness of this relative price measure. There also appears to be marginally significant residual autocorrelation in these equations.

In Panel B of Table 4, equations (ii) and (iv) report analogous results for high-powered money as the measure of external money with general features that are again broadly consistent with other studies. Under our theory, real bank reserves should enter negatively in such price level equations if high-powered money is the controlled measure of external money. This is borne out by significant negative coefficients in both the unconstrained equation (ii) and constrained equation (iv). As before, the service charge variable has a tendency to change sign when the unit constraint on high powered money is imposed.

Overall, the results of Table 4 are broadly consistent with the theoretical stories told in the sections above. The negative influence of real reserves on the price level potentially is important, both in terms of explaining postwar price-level behavior and in explaining the apparently anomalous behavior of the price level during the interwar period. Finally, additional work needs to be done in producing measures of the market prices of bank services.

IV. Conclusions

This paper describes a class of real business cycle models that is capable of accounting for the relation between money, inflation, and economic activity, providing a coherent alternative framework to the monetary theories of the business cycle advanced by Lucas (1973) and Fischer (1977). Although the empirical work presented is simplistic, we draw two main lessons from it. First, much of the contemporaneous correlation of economic activity and money is apparently with inside money, with inflation principally resulting from changes in the stock of fiat (or outside) money and variations in real activity. This empirical observation implies that care should be taken in empirical studies to distinguish inside from outside money. Second, future work along these lines may have to consider policy responses that are broad enough to produce variations in outside money that are correlated with real activity.

A main direction of our future work in this area will be to develop the implications of the analysis for security returns so that the general equilibrium predictions for these variables can be exploited in tests of the model. This topic is especially important because Sims (1980) and Fama (1981) have provided some hints about the interrelationship of money, asset returns, and real activity.

In conclusion, it seems worthwhile to discuss two recurrent comments on this line of research that we have received. First, there has been a surprising willingness on the part of many individuals to simultaneously argue that our model (a real business cycle model with an explicit banking sector and central bank) is probably observationally equivalent to many existing monetary theories *and* that a "common sense" view leads one to prefer alternative models as descriptions of reality.[15] This line of argument puzzles us, since it was presumably on empirical grounds (not common sense) that the profession rejected pre-Keynesian "equilibrium theories" of the business cycle that stressed real causes of economic fluctuations.

Second, some individuals have argued that market failure is central to both the understanding of cyclical fluctuations and the primary reason for economists to study these phenomena. Our view is that widespread market failure need not be a necessary component of a theory of business fluctuations, and that real equilibrium business cycle theory promises to make important contributions to positive economics. This perspective, however, is not inconsistent with the view

[15] Grossman (1982) makes an explicit statement of this view.

that the accumulation of scientific knowledge may lead to the design of more desirable government policies toward business fluctuations (such as tax and expenditure policies) or toward the regulation of the financial sector.

REFERENCES

Azariadis, Costas, "Escalator Clauses and the Allocation of Cyclical Risks," *Journal of Economic Theory*, June 1978, *18*, 119–55.

Barro, Robert, (1981a) "Unanticipated Money Growth and Economic Activity in the U.S.," *Money, Expectations, and Business Cycles*, New York: Academic Press, 1981, ch. 5.

______, (1981b) "Output Effects of Government Purchases," *Journal of Political Economy*, December 1981, *89*, 1086–121.

______, "Long-Term Contracting, Sticky Prices and Monetary Policy," *Journal of Monetary Economics*, July 1977, *3*, 305–16.

Black, Fischer, "Active and Passive Monetary Policy in a Neo-Classical Model," *Journal of Finance*, September 1972, *27*, 801–14.

Boschen, John and Grossman, Herschel, "Tests of Equilibrium Macroeconomics Using Contemporaneous Monetary Information," *Journal of Monetary Economics*, November 1982, *10*, 309–34.

Burns, Arthur A. and Mitchell, Wesley, *Measuring Business Cycles*, New York: National Bureau of Economic Research, 1946.

Fama, Eugene, "Banking in the Theory of Finance," *Journal of Monetary Economics*, January 1980, *6*, 39–57.

______, "Stock Returns, Real Activity, Inflation and Money," *American Economic Review*, September 1981, *71*, 545–65.

______, "Inflation, Output and Money," *Journal of Business*, April 1982, *55*, 201–31.

Fischer, Stanley, "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Role," *Journal of Political Economy*, February 1977, *85*, 191–205.

______, "A Framework for Monetary and Banking Analysis," Working Paper, Massachusetts Institute of Technology, July 1982.

Friedman, Benjamin, "The Relative Stability of Money and Credit Velocities in the United States: An Overview of the Evidence," Working Paper, Harvard University, November 1981.

Friedman, Milton and Schwartz, Anna, *A Monetary History of the United States*, Princeton: National Bureau of Economic Research, Princeton University Press, 1963.

Grossman, Herschel, Review of James Tobin's *Asset Accumulation and Economic Activity*, in *Journal of Monetary Economics*, July 1982, *10*, 134–38.

Hamilton, James, "Oil and the Macroeconomy Since World War II," *Journal of Political Economy*, April 1983, *91*, 228–48.

Hodrick, Robert and Prescott, Edward, "Post-War U.S. Business Cycles: An Empirical Investigation," Working Paper, Carnegie-Mellon University, November 1980.

King, Robert, "Monetary Information and Monetary Neutrality," *Journal of Monetary Economics*, March 1981, *7*, 195–206.

Kydland, Fynn and Prescott, Edward, "Time to Build and Aggregate Fluctuations," *Econometrica*, November 1982, *50*, 1345–70.

Lilien, David, "Sectoral Shifts and Cyclical Unemployment," *Journal of Political Economy*, August 1982, *90*, 777–93.

Long, John and Plosser, Charles, "Real Business Cycles," *Journal of Political Economy*, February 1983, *91*, 39–69.

Lucas, Robert, "Some International Evidence on Output-Inflation Tradeoffs," *American Economic Review*, June 1973, *63*, 326–34.

______, "Understanding Business Cycles," in K. Brunner and A. Meltzer, eds., *Stabilization of the Domestic and International Economy*, Vol. 5, Carnegie-Rochester Series on Public Policy, *Journal of Monetary Economics*, Suppl. 1977, 7–29.

______, "Asset Prices in an Exchange Economy," *Econometrica*, December 1978, *46*, 1429–48.

______, "Methods and Problems in Business Cycle Theory," *Journal of Money, Credit and Banking*, November 1980, *12*, 696–715.

Mitchell, Wesley, *Business Cycles: The Problem and Its Setting*, New York: National Bureau of Economic Research, 1930.

Nelson, Charles and Plosser, Charles, "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications," *Journal of Monetary Economics*, September 1982, *10*, 139–62.

Sargent, Thomas and Wallace, Neil, "Rational Expectations, the Optimal Monetary Instrument and the Optimal Money Supply Rule," *Journal of Political Economy*, April 1975, *83*, 241–54.

Sims, Christopher, "Money, Income, and Causality," *American Economic Review*, September 1972, *62*, 540–52.

______, "Comparison of Interwar and Postwar Business Cycles: Monetarism Reconsidered," *American Economic Review Proceedings*, May 1980, *70*, 250–57.

Stockman, Alan, "Anticipated Inflation and the Capital Stock in a Cash-in-Advance Economy," *Journal of Monetary Economics*, November 1981, *8*, 387–94.

Tobin, James, "Commercial Banks as Creators of 'Money'," 1963; reprinted in *Essays in Economics*, Vol. 1: *Macroeconomics*, North-Holland: Amsterdam, 1971, ch. 16.

______, "Money and Economic Growth," 1965; reprinted in *Essays in Economics*, Vol. 1: *Macroeconomics*, North-Holland: Amsterdam, 1971, ch. 9.

______, "Money and Income: Post Hoc Ergo Propter Hoc?," 1970; reprinted in *Essays in Economics*, Vol. 1: *Macroeconomics*, North-Holland: Amsterdam, 1971, ch. 24.

Board of Governors of the Federal Reserve System, *Annual Statistical Digest*, Washington, various issues.

______, *Banking and Monetary Statistics, 1941–1970*, Washington, 1976.

U.S. Department of Commerce, *Business Statistics, 1979*, Washington, 1979.

______, *National Income and Product Accounts of the United States, 1929–1974*, Washington.

Internal Bargaining, Labor Contracts, and a Marshallian Theory of the Firm

By HAJIME MIYAZAKI*

The theme of this paper is due to Alfred Marshall who states that "nearly the whole income of a business may be regarded as a ... *composite quasi-rent* divisible among different persons in the business by bargaining supplemented by custom and by notion of fairness...[And] the division of quasi-rents entails *de facto* some sort of profit-loss sharing between almost every business and its employees" (*Principles of Economics*, 8th ed., Book VI, ch. VIII, Sec. 10, pp. 520–21). Marshall emphasizes that a significant portion of composite quasi rent is derived from the organization of business and would be lost if employer-employee connections were dissolved. In this view, bargaining can even be tacit insofar as there is an organizational basis which ensures agreement among the concerned parties. Consequently, internal bargaining models of the firm may be applicable to the analysis of European-style codetermination and Japanese-style labor management. Also, as the relative bargaining strength between labor and management changes, such a view of the firm admits as its extrema, the organization of a *labor-managed firm* (*LMF*), and the textbook case of a *profit-maximizing firm* (*PMF*).

Masahiko Aoki (1980, 1982) and Jan Svejnar (1982) have recently explored models of an internal bargaining firm. Their analyses, however, consider the firm's long-run behavior only under restricted circumstances without market uncertainty. The relevance of the internal bargaining approach to profit sharing and labor management would be enhanced if we integrated the firm's short-run wage-employment policies with its long-run plans for the rate of growth and capital-labor substitutions. This I do by underscoring the firm's organizational basis, which is essentially a long-term contractual association between workers and management. I combine Costas Azariadis' (1975) and Martin Baily's (1974) apparatus of labor contracting with efficient bargaining to investigate the effect of varying degrees of labor's bargaining power upon the firm's short- and long-run policies under uncertainty. These results are then contrasted with the empirical findings of collective bargaining (for example, Richard Freeman and James Medoff, 1981), conventional results of *LMF* models (Jaroslav Vanek (1970); Benjamin Ward, 1958), and the recent macroeconomic approaches to wage-employment adjustments (Ian McDonald and Robert Solow, 1981).

I. Scenario

My discussion will be limited to the case of two organizational persons: "labor" and "management." The two-person setup assumes that all workers are homogeneous, and that management makes decisions that reflect unanimous approval by the firm's investors.[1] The major hypotheses that are maintained throughout the paper are as follows:

HYPOTHESIS 1: *Labor's objective is to maximize the worker's expected utility.*

HYPOTHESIS 2: *Management's objective is to maximize expected profits given bargaining constraints.*

*Department of Economics, Stanford University, Stanford, CA 94305. I thank Masahiko Aoki, Haruo Imai, Masahiro Okuno, and Kotaro Suzumura for valuable discussions. In writing the present version, I have benefited from suggestions of Paul Evans, Thomas MaCurdy, Ian McDonald, Hugh Neary, William Novshek, John Pencavel, Craig Riddell, and David Weisberg. I am solely responsible for any remaining errors and omissions. Research support from the Stanford Workshop on Factor Markets is gratefully acknowledged.

[1] An alternative interpretation is to regard management as an arbitrator who synthesizes conflicting demands of shareholders and workers.

HYPOTHESIS 3: *Work hours are institutionally fixed; the number of workers employed represents labor input.*

HYPOTHESIS 4: *The firm sells its output in a competitive market, and uncertainty exists as to the output price.*

When the Marshallian quasi rent is primarily due to firm-specific organizational resources, the assumption of a competitive output market is not incongruous to the framework of internal bargaining.[2] The distinguishing feature of the contractual framework is the supposition of a labor pool that is attached to the firm for the duration of labor contracts. Without such a labor pool, the meaning of short-run furlough becomes opaque. More importantly, the rent accrues to, and (partly) because of, the firm-specific labor pool.

I develop the analysis in two stages. In Sections II–IV, a basic model is discussed in which the firm's capital input level is fixed and bargaining focuses on the firm's wage-employment policy and the contractual labor pool size. In Section VI, I extend the basic model so that labor and capital become smoothly substitutable variables directly subject to the labor vs. management bargaining. Additionally, I let bargaining determine the long-run growth rate of the firm. Interestingly, the important economic features of the full-fledged long-run model are almost all captured by the basic model. Note also that the assumption of a fixed capital input has been regularly deployed within the contractual framework in discussing the long-term aspects of the *PMF*'s employment policies (for example, Baily; Sanford Grossman and Oliver Hart, 1981; Bengt Holmstrom, 1983; Sherwin Rosen, 1983). In this regard, I derive from the basic model new comparative statics relating bargaining variables to the labor pool size and layoff frequencies. To avoid confusion in the course of our development, I use "long term" to denote the market duration in which the labor pool size changes but the capital input is fixed as in the basic model; I reserve "long run" for the case in which the capital input as well as the labor pool size changes as in the model of Section VI.

II. A Basic Model

It is assumed here that the firm's capital input level is fixed. Suppressing the fixed capital, I write the firm's production function as f with the following properties in labor inputs: $f'>0$, $f''>0$, $f'(0)=\infty$, and $f'(\infty)=0$. The competitive output price s is a random state of nature to the firm. The bargainers negotiate a state contingent contract (*ex ante*) knowing only the probability distribution of s. The agreement will then be carried out when the state becomes known (*ex post*). The two-period model is interpreted as a stylized long term under a stationary-state perception: *ex ante* represents a tunnel vision of the long term over the two periods whereas the short run corresponds to a particular *ex post* state. *Ex ante* contractual bargaining settles the conduct of production and rental division for each short run in accordance with the bargainers' expectations over all possible short-run states. State-contingent terms of a contract ipso facto define the firm's short-run behavior as it responds to various s.[3] Thus, the two-period setup allows the standard Marshallian interpretation of short and long term.

In accordance with the notion of a long-term contract, the size of a firm-specific labor pool N must be chosen *ex ante* and remain fixed *ex post*. It is then stipulated that $n(s)$, the workforce actually employed in s, must come from within the labor pool, $n(s)\le N$ for all s; there may be temporary layoffs of $N-n(s)$ workers in some s.[4] Besides $n(s)$, state-contingent terms are $w(s)$, the wage payment made to an employed worker in s, and $c(s)$, the compensation paid by the firm

[2] For example, Marshall's interest was focused more on the competitive firm's composite rent than the pure monopoly rent in his Book VI, ch. III, Sec. 9–10.

[3] For example, $y(s)$, the firm's output rate contingent upon s, traces the conventional Marshallian short-run supply curve as s takes on different values.

[4] This labor pool constraint together with the fixed length of workday is a standard assumption in the labor contract literature beginning with Azariadis and Baily.

to a laid-off worker in s. The across-state profiles of $w(s)$, $n(s)$, and $c(s)$ sum up the firm's short-run response to different s in a manner that is consistent with long-term expectations.

A. *Management*

For each state I define

$$(1) \qquad \pi(s) \equiv sf(n(s)) - w(s)n(s) - c(s)[N - n(s)],$$

so that the expected return on capital is simply $E\pi(s)$ where E is taken across s. Management is constrained by $E\pi(s) \geqq R^*$, R^* being the normal return on capital available elsewhere in the economy. Management's maximand is therefore $E\pi(s) - R^*$.

B. *Labor*

The worker's utility $u(\cdot)$ is defined on wage income.[5] It is assumed that $u' > 0$, $u'' < 0$, $u'(0) = \infty$, and $u'(\infty) = 0$. The maximum utility value of home production is assumed to be state invariant and given by $u(k)$; this k is the worker's short-run reservation wage income.[6] With the firm's own layoff compensation $c(s)$, the worker's *mean utility* $v(s)$ in s is given by

$$(2) \qquad v(s) \equiv [n(s)/N]u(w(s)) + \{1 - [n(s)/N]\}u(c(s) + k).$$

The worker's *expected utility* is $Ev(s)$. The worker's long-term opportunity cost of joining the firm's labor pool is measured by V^*, the expected utility value of working elsewhere. Labor's maximand is $Ev(s) - V^*$.

C. *Bargaining-Theoretic Aspects*

Labor and management bargain for a contract specifying N and $s \rightarrow [n(s), w(s), c(s)]$. The firm's production function, worker's utility, and the probability distribution of s jointly determine the feasible set, a collection of $(Ev, E\pi)$ pairs which are attainable by contracts. The set of maximal feasible pairs defines the usual Paretian contract curve for the two. A contract is said to be *efficient* if it attains $(Ev, E\pi)$ on the contract curve. For every *disagreement point*, (V^*, R^*), a bargaining set $B(V^*, R^*)$ is a set of feasible pairs such that $(Ev, E\pi) \geqq (V^*, R^*)$. The intersection of the contract curve and the bargaining set is the *bargaining frontier*. I consider only those bargaining sets whose frontiers are nondegenerate, and for which $(V^*, R^*) \gg (k, 0)$. Bargaining is said to be efficient if it results in a contract on the bargaining frontier.[7] The market environment and negotiable options are summarized by the shape of bargaining sets. Relative bargaining power may be related to the shape of the bargaining set, but it is also affected by exogenous circumstances such as social customs, notions of fairness, and the bargainer's noneconomic power.[8]

It can be shown that the set of feasible contracts is always convex if N is exogenous to bargaining: N may be considered to be fixed if the bargaining horizon is short. Without institutional constraints, however, N will adjust as changes occur in the firm's long-term bargaining environment. Thus, in this scenario, N is determined by bargaining over long-term contracts. In other words, the long-term feasible set is equivalent to the union of all short-run feasible sets indexed by N. But this union fails to preserve convexity for a large class of labor-management

[5] Since my analysis is partial equilibrial, I consider that s, w, c, and k are expressed in real terms deflated by a general price index of each state.

[6] The term k is the maximal value of consumption with home production including the government dole. It is also possible to interpret k as the perceived wage income earnable on a temporary job outside the worker's own contract firm.

[7] Labor's strikes and management's lockouts are ruled out under the efficiency rationale with a fixed disagreement point.

[8] See, for example, Marshall. In game theory, each solution concept incorporates relative bargaining power differently. In Nash solution, it is exogenous to a bargaining set. In Raiffa solution, it is influenced by the shape of a bargaining set.

conflict.[9] Technically, nonconvexities in bargaining sets can be filled by allowing probability mixtures on outcomes of negotiation. This means a randomization of N in this case, but such convexification has little practical support in most empirical settings of labor-management bargains.[10] Summing up this case, the Paretian contract curve is not usually concave; hence bargaining sets are not necessarily convex. The cause of the nonconvexity lies in the long-term nature of bargained contracts. The implication of such nonconvexity is as follows.

Because game-theoretic solution concepts of bargaining usually require bargaining sets to be convex, we do not have formal recourse to most bargaining solutions.[11] Note, however, that even if the bargaining sets were convex, we would not have a priori criterion to select a solution concept from several competing theories of bargaining. Since each solution concept has incorporated a different axiomatic quantification of bargaining power, to obtain a robust result we would have to calculate outcomes for each solution concept and glean some generalities from them (for example, W. Craig Riddell). Frequently in those exercises, pertinent characteristics of bargained contracts were derived from the efficiency and uniqueness of the bargaining outcome, combined with intuitive notions of division rules. The foregoing observation suggests that the following approach is sensible for our purpose.

As labor and management negotiate the firm's policies under various circumstances, I postulate that the mode of bargaining meets the following regularity conditions on all bargaining sets, convex or nonconvex.[12]

Condition 1: Bargaining entails a unique efficient outcome on each bargaining set.

Condition 2: On a given bargaining set, the rental-share ratio, $(Ev-V^*)/(E\pi-R^*)$, in the outcome, is positively related to labor's relative bargaining power vis-à-vis management. That is, a stronger bargainer should claim, *ceteris paribus*, a larger share of the rent.

Under these conditions we may make careful deductions about the effect of changes in underlying bargaining parameters on the firm's economic decisions by inspecting how a bargaining set shifts. We may also translate various bargaining outcomes into the comparative statics of efficient contracts. A comparative statics exercise on N, for example, can be viewed as a way to compare long-term outcomes between firms with different bargaining conditions.

III. Efficient Contracts

To relate efficient bargaining outcomes to contracts, it is very useful first to study two extreme cases, *PMF* and *LMF*, from which the behavior of a more realistic firm can be deduced as an intermediate case. A *PMF* is a firm in which all the rent accrues to management; it maximizes $E\pi$ subject to the expected utility constraint. It is an Azariadis-Baily contractual firm modified by the firm's *own* provision of layoff compensation. The opposite extreme is an *LMF* wherein egalitarian labor maximizes the individual member's Ev subject to the expected return constraint. Mathematically these two programs are dual to each other. An efficient contract that achieves (V,R) can be characterized

[9] A detailed explanation and an example of nonconvex bargaining sets within the specifications of my model are available upon request. In McDonald and Solow, and W. Craig Riddell (1981), N is fixed and the convexity follows immediately. Aoki (1980) recognizes such nonconvexity; his appeal to Zeuthen-Harsanyi process merely subsumes, but does not exorcise the nonconvexity. Svejnar does not explicitly deal with the issue.

[10] McDonald and Solow also express reservations about convexification by randomized outcomes.

[11] In particular, it is not appropriate for us to rely on a Nash-solution concept. An asymmetric Nash-bargaining solution would have solved the following program: $\mathrm{Max}(Ev-V^*)^{\alpha}(E\pi-R^*)^{1-\alpha}$ subject to the usual constraints where α is labor's relative bargaining power $(0\leqq\alpha\leqq1)$. It could be axiomatically justified if bargaining sets were convex (see E. Kalai, 1977). This specification has been most frequently used in applied bargaining theory primarily because it is most palatable to marginal calculus. For example, Aoki, McDonald and Solow, and Riddell use it with $\alpha=1/2$; Svejnar considers an n-person case: $\alpha_1+\ldots+\alpha_n=1$.

[12] In fortuitous circumstances in which bargaining sets are convex, both Nash and Raiffa solutions, for example, meet these two conditions.

equivalently as a solution to a *PMF* program subject to $Ev \geqq V$ or as a solution to an *LMF* program subject to $E\pi \geqq R$.[13] In either characterization the following lemmas obtain.

LEMMA 1: *An efficient contract satisfies the following conditions*:

(i) $[N-n(s)][sf'(n(s))-k]=0$; *that is*, $n(s)=N$ *iff* $k \leqq sf'(N)$,

(ii) $w(s)=c(s)+k$,

(iii) $w(s)=w$ *for all* s; $c(s)=c$ *for all* s *such that* $n(s)<N$; *and* $c(s)=0$ *whenever* $n(s)=N$,

(iv) $Esf'(n(s))=w$.

LEMMA 2: *To each outcome of efficient bargaining, there corresponds a unique efficient contract.*

Although the lemmas can be readily derived from the first-order conditions of either program, I give here an elementary explanation that relies on efficient risk sharing. By convention, it is presented from the viewpoint of the *PMF*, but its translation into the *LMF* program is straightforward. I first prove conditions (ii) and (iii), from which the other properties immediately follow.

To see condition (ii), suppose that the worker faces in s an uncertain prospect of income loss due to layoffs. The risk-averse worker is then willing to insure fully against the income risk by paying premia out of his wage income so that he can be indemnified if laid off. A noteworthy point is that such a full insurance scheme can be devised for the N workers without changing either $n(s)$ or the total labor bill paid in s.[14] Full coverage means that the worker has increased $v(s)$ by equalizing the marginal utility of income between two random outcomes, employment or layoffs, in s; that is, $w(s)=c(s)+k$. This full insurance scheme increases $v(s)$ without decreasing $\pi(s)$; a Pareto improvement is possible if the bargainers agree on a fully insuring wage-employment package. Since the scheme relies only on mutual risk sharing by pooling of wage payments among N workers, it can be implemented by the *LMF* or the labor union. It is a self-insurance that requires no outside capital markets. Although the self-insurance scheme eliminates income risk within a given state, the worker still faces income uncertainty across s. The risk-neutral firm can then act as an insurance agent to stabilize the worker's marginal utility of income across states, and (iii) follows.[15] By doing this, management succeeds in increasing $E\pi(s)$ because the average labor bill will be subsequently reduced as labor pays implicit premia for the across-state income stability.[16]

In view of conditions (ii) and (iii), an efficient contract can be represented simply as $[w, c, n(s), N]$. Much convenience will be gained by redefining $\pi(s)$ and $v(s)$ for the set of efficient contracts as follows.

$$\pi(s)=sf(n(s))-kn(s)-cN, \tag{3}$$

$$v(s)=u(c+k) \quad \text{for all } s. \tag{4}$$

[13] This qualitative equivalence in the necessary conditions of an efficient contract does not mean that the effects upon N of comparative static changes are the same between the *PMF* and *LMF*. Indeed, a thrust of my analysis is to show how they differ.

[14] Suppose $w(s)>c(s)=k$ and $n(s)<N$ in s, and consider an insurance policy which charges a premium rate of $1-(n(s)/N)$ for the full coverage of income loss. This policy breaks even because the total premia collected from N workers is exhausted by the total indemnity payments to $N-n(s)$ laid-off workers. Define $\overline{w}(s)=(n(s)/N)w(s)+(1-(n(s)/N))(c(s)+k)$. With the above insurance, the workers' utility in s is $u(\overline{w}(s))$ regardless of employment status, and $u(\overline{w}(s)) > (n(s)/N)U(w(s))+(1-(n(s)/N))u(c(s)+k)$. This scheme can be implemented by a contract which employs $n(s)$, pays $\overline{w}(s)$ to the employed and $\overline{w}(s)-k$ as compensation to the laid off; the total labor bill in s is unchanged at $n(s)w(s)+(N-n(s))c(s)$. The foregoing logic is applicable even if $w(s)<c(s)+k$ to begin with. Charge the premium rate of $n(s)/N$ for the full coverage of $c(s)+k-w(s)$.

[15] Since $w(s)=c(s)+k$, the average labor bill across s is $(NEw(s))-k(N-En(s))$. By the worker's risk aversion, $Eu(w(s))<u(Ew(s))$. Hence, there exists w^* such that $w^*<Ew(s)$ and $Eu(w(s))<u(w^*)$. With the same $n(s)$ and N, but with a new w^*, the labor bill is reduced and $E\pi$ is increased without lowering Ev.

[16] $E\pi(s) \geqq R$ implies that the *LMF* has access to the *ex ante* perfect capital market. It enables the *LMF* to make state-contingent debt repayment plans as $\pi(s) \geqq R(s)$ where $R(s)$ adjusts with s. In a socialist labor-managed economy, the central bank can, in theory, achieve this by acting as the ultimate insurer against state-of-nature uncertainty.

Equation (3) says that with an efficient contract, the firm's short-run cost consists of the quasi-fixed cost (cN) and variable cost $(kn(s))$.

The efficient *PMF* and *LMF* programs thus become

$$
\begin{aligned}
&\text{(P) Max } E\pi(s) && \text{subject to } n(s) \leqq N \\
& && \text{and} \quad u(c+k) \geqq V; \\
&\text{(L) Max } u(c+k) && \text{subject to } n(s) \leqq N \\
& && \text{and} \quad E\pi(s) \geqq R.
\end{aligned}
$$

It is straightforward to confirm that a unique solution exists for each program. Hence, Lemma 2 follows.

From (P), the first-order condition on $n(s)$ can be expressed as

$$(5) \qquad \gamma(s) = sf'(n(s)) - k,$$

where $\gamma(s)$ is a probability-adjusted nonnegative Lagrangian multiplier for the labor pool constraint in s. If $\gamma(s) > 0$, then $n(s) = N$, and $n(s) < N$ only if $\gamma(s) = 0$. Hence, layoffs occur in s if and only if $sf'(n) = k$ for some n strictly less than N. This employment rule is valid once N is fixed; it stipulates *ex post* efficient use of the labor pool. As such it is a variant of the Robert Hall-David Lilien (1979) rule. Finally, the condition on N is given by $E\gamma(s) = c$, which is combined with (5) to give condition (*iv*).

IV. Analysis

Conditions (*i*)–(*ii*) of Lemma 1 say that the efficient use of a given labor pool entails the marginal productivity rule taking k as the short-run shadow wage rate.[17] These conditions imply the firm's short-run behavior as in the following lemma.

LEMMA 3: *Let s^* be a short-run benchmark state defined by $s^*f'(N) = k$. Then, $n(s)$ is increasing on $(0, s^*)$ and constant at N on $[s^*, \infty)$. Consequently, both the output supply and employment level of an efficient, internal-bargaining firm move procyclically with s.*

An interesting application of Lemma 3 is to a firm with strong labor or *LMF*. Typically the latter has been criticized for its perverse tendency to have a negatively sloped output supply curve in the short run, especially when labor is the only variable input as in Ward and Vanek. Yet, Lemma 1 says unambiguously that if egalitarian utility maximization is applied to all N members, the short-run supply curve of outputs ought to have a normal shape even for a purely labor-managed firm. This is because efficient use of the labor pool is essential to any rent-conscious organization.

We can also use Lemma 1 to link the long-term aspects of bargaining to the comparative statics of changes in N.

LEMMA 4: *In the PMF program* (P), *the solution contract is a function of V and has the following properties*:

(*i*) *c is increasing in V*,
(*ii*) *N is decreasing in V*,
(*iii*) *$n(s)/N$ is increasing in V for each s.*

My proof makes use of the $Ev \geqq V$ constraint, which always binds in the solution. From $u(c+k) = V$, a rise in V necessitates an increase in c. Since $\gamma(s) = sf'(n(s)) - k$ still holds, but $Esf'(n(s))$ must equate with a now higher $c + k$, N must decrease. Consequently, the benchmark s^* decreases and $n(s)/N$ increases in every state where layoffs previously occurred. The corresponding *LMF* lemma can be similarly derived.[18]

Bargaining Conditions 1 and 2 imply that stronger labor, *ceteris paribus*, obtains a contract promising a higher value of Ev on a given bargaining set. Setting the bargaining outcome as $Ev = V$, we can restate Lemma 3

[17] Condition (*i*) of Lemma 1 depends only on the self-insurance scheme within s. The availability of across-state insurance policies, however, affects N, and the actual short-run shut-down price.

[18] In (L), c is decreasing in R, N is increasing in R, and $n(s)/N$ is decreasing in R for each s. Since the *LMF* maximizes income per capita, $N \to 1$ as $R \to 0$. With a strictly concave f, a nondegenerate *LMF* is formed only to share R the burden of capital cost payments.

as follows:

PROPOSITION 1: *Strong labor succeeds in raising Ev by a combination of higher wage-compensation payments and lower probabilities of layoffs in all states, via a reduction in N. Conversely, stronger management earns a higher return on capital by being able to pay a lower wage rate and by enlarging the labor pool, resulting in higher layoff probabilities within and across states.*

The results thus far derived could be connected with the effect of different disagreement points provided, for instance, that a *ceteris paribus* rise in one party's disagreement position is unambiguously related to an increase in that party's relative bargaining power. When a parametric change causes a shift in the contract curve, the relationship between relative bargaining power and the comparative static changes in the bargaining outcome is not easily predictable. In the remainder of this section, I report somewhat novel results on the effects of changes in k and of shifts in the state distributions, both of which affect the shape and position of the bargaining frontier.

A. *Effects of k*

A shift in either the value of activities outside the firm or the amount of government-financed dole can change k. Because k is the short-run shadow wage rate, a change in k immediately affects $n(s)$ in each s. Then, the bargainers will take into account such shifts in the utilization of N, and make a different agreement on N itself. The main result is that a rise in k expands the contract curve, and also tends to enlarge N.

Consider a *PMF* and let k increase by Δk *ceteris paribus*. Because management compensates labor exactly at V, the firm can lower c by \$1 for each \$1 increase in k. Such a matching reduction in c immediately enables the firm to capture the net saving in labor cost by $(N-n(s))(\Delta k)$ in each s. The firm can further improve profits by reducing $n(s)$ to equate $sf'(\cdot)$ to $k+\Delta k$. Consequently, unless the firm increases N, $Esf'(\cdot)$ exceeds $c+k=u^{-1}(V)$; a further increase in $E\pi$ will be garnered by enlarging N. In other words, the reduction in c due to a rise in k allows the firm to expand N without increasing cN; this allows $\pi(s)$ to increase by $\gamma(s)$ whenever the labor pool constraint was previously binding. The above inspection can be made rigorous by using (5) and (*iv*) of Lemma 1 to confirm that $\text{sgn}[\partial N/\partial k]_V$ is positive. Finally, note that a higher k raises the benchmark state s^* and that layoff frequencies rise both within and across states.[19]

The logic behind the *PMF*'s response to an increased k is equally valid in the case of an *LMF*. For the *LMF*, however, any potential gain in $E\pi$ will be translated into a higher per worker remuneration $c+k$, so that the return constraint will once again be met exactly. To the extent that the *LMF* raises $c+k$ and prefers lower probabilities of layoffs, the propensity of the labor pool expansion will therefore be less than that of the *PMF*.[20]

We have seen that a rise in k permits a reduction in c without decreasing $c+k$, and that a reduction in c allows a larger N which in turn increases $E\pi$ by $E\gamma(s)$. From this observation one infers that, for firms intermediate to *PMF* and *LMF*, N increases with k as long as the expanded contract curve enables a Pareto improvement in the bargaining outcome for both parties. Such a condition is likely to be satisfied when k rises in a small increment. When k increases in a sufficiently large magnitude, however, depending on the shape of the expanded bargaining set, it is possible that N decreases in the eventual bargaining outcome. For a

[19] The state s^* is given by $s^*f'(N)=k$. It follows that $\partial s^*/\partial k=\{f'(N)-kf''(N)[\partial N/\partial k]_V\}/f'(N)>0$ because $[\partial N/\partial k]_V=-\text{Prob}\{0\leqq s\leqq s^*\}/f''(N)(Es)>0$.

[20] The sense in which the *LMF* has less incentive to expand N is stated more precisely as follows. Given k, pick an arbitrary (V,R) on the contract curve. It can be viewed as a solution either to (P) or (L) with V and R in respective constraints. Let subscripts *LMF* and *PMF* represent the respective firm's comparative static responses. It can be shown $[\partial N/\partial k]_{LMF}=[\partial N/\partial k]_{PMF}-E(N-n(s))[\partial N/\partial R]_{LMF}$. We can then verify $[\partial N/\partial k]_{LMF}>0$, $[\partial N/\partial R]_{LMF}>0$, and $[\partial N/\partial k]_{PMF}>[\partial N/\partial k]_{LMF}>0$.

larger k drastically strengthens labor's bargaining power to the extent that management's absolute share in the rent declines.

B. *Increasing Uncertainty*

What will be the effect of a mean-preserving spread (MPS) in the state distribution upon N and the contract curve? Once again, the consequences for the PMF and LMF are spelled out separately, to infer the result of the intermediate case. Let us start with a PMF. Define $M(s|N)=\text{Max}\{sf(n)-kn-cN|n\leqq N\}$ for a fixed N and given s. Note that $\text{Max}\,E\pi(s)=\text{Max}_N EM(s|N)$ for the PMF. Since $M(s|N)$ is a convex function of s, and because $n(s)$ is adjusted contingent on s, the increased variability of s induces a higher $E\pi(s)$ *ceteris paribus*.[21] It implies that the bargaining frontier expands after the MPS. The effect on N can be gleaned from condition (*iv*) of Lemma 1, which I now write as $Esf'(n(s|N))=c+k$ to emphasize $n(s|N)=\arg\text{Max}\,M(s|N)$. Since $sf'(n(s|N))$ is also a convex function of s, a MPS on s increases $Esf'(n(s|N))$. From the second-order maximum condition for N, N must then expand in order to restore the condition $Esf'(n(s|N))=c+k$ given $u(c+k)=V$.[22]

Similarly, it can be shown that the LMF's $u(c+k)$ increases as the result of MPS. As to the LMF's membership size N, however, MPS has an ambiguous effect. This can be seen from rewriting (*iv*) of Lemma 1 for the LMF as

$$(6)\quad Esf'(n(s|N)) = \frac{Esf(n(s|N))-kn(s|N))-R}{N}+k,$$

which comes from solving $E\pi(s|N)=R$ for c. Once again both $sf'(n(s|N))$ and $M(s|N)$ are convex functions of s when N is fixed. Depending on MPS and the curvature of f, either side can increase faster than the other in (6). If MPS makes the right-hand side larger than the left-hand side, then an LMF will find it Ev-maximizing to reduce N; conversely, N will increase under MPS if the left-hand side becomes larger than the right-hand side. Either example can be constructed. In general, the direction of a change in N in response to increased uncertainty is a priori ambiguous. Once again, the complication is due to labor's incentive to increase $c+k$.

Because the short-run workforce can be adjusted contingent on s, increased uncertainty actually expands the bargaining frontier. These MPS exercises confirm that there are changes in the firm's environment which can induce a long-term adjustment in the opposite direction, depending on the internal distribution of bargaining power. Thus, we cannot make a systematic prediction on the MPS effect on N with respect to labor's bargaining power unless we know a great deal about how the bargaining set shifts and how the bargaining process works.

V. Qualifications

In the basic model, short-run productive efficiency was facilitated by the optimal compensation payments to those laid off.[23] When firms do not provide their own layoff compensations as in Azariadis, Baily, Holmstrom, and McDonald-Solow, there will be short-run overemployment of workers. Overemploy-

[21] This property of the "profit function" was first noted by Walter Oi (1961). This and other consequences of MPS in this section readily follow from a well-known statistical inequality of a convex function. The key is that $n(s)$ is contingent on the random variable s. See Larry Epstein (1978) for a general treatment of such MPS uncertainty.

[22] My analysis here resembles Riddell's investigation of bargaining under uncertainty. I, however, differ from him both in results and structure on two grounds. First, N itself is endogenous to my model whereas it is assumed given in his analysis. Thus, contrary to his adaptation of the Azariadis-Baily-type PMF firm, the wage rate $w(s)$ in my model becomes sensitive to an increase in uncertainty. Second, I differ from him in the treatment of the disagreement point.

[23] The efficiency role of $c(s)$ has been recognized by Grossman and Hart, Rosen, and exploited by myself and Hugh Neary (1983) in solving a classical conundrum of the LMF.

ment occurs because the employment adjustment becomes the only means to mitigate partially the *ex post* income risks. Realistically it can be argued that a suboptimal provision of $c(s)$ is a second best norm in a world with imperfect information and enforceability problems. Even a complete absence of the compensation scheme (i.e., $c(s)=0$), however, would not alter the main conclusions obtained in the basic model (see Holmstrom also).

Though the McDonald-Solow analysis is concerned with short-run bargaining (N fixed), my results on N are relevant to their macroeconomic question regarding the workings of primary and unionized labor markets. This is because these labor market sectors can be identified with a collection of contract-based employment exchanges. For this, I interpret the two-period as a "medium-term," say three years, which.is a rule-of-thumb average duration of union-bargained contracts in the United States. During the contract duration, the contingent terms are carried out as the firm experiences short-run business fluctuations. In reissuing contracts, the firm will decide on its regular labor force depending on the long-term outlook. It is evident that the forces affecting labor pool size generate macroeconomic consequences especially in the staggered setting of bargaining across firms.

I emphasize, however, that temporary layoffs (and even separations from the labor pool) in the bargaining regime can be done efficiently.[24] Labor's risk aversion and bargaining efficiency predict sticky short-run wages, but the bargained wage path itself is decoupled from the actual implementation of the short-run marginal productivity rule. Even in the absence of an optimal compensation scheme, it is difficult to find Keynesian-type short-run *under*employment in the internal-bargaining regime. Still, to the extent that the firm's adjustment in the labor pool is influenced by the conditions of internal bargaining, *long-run* implications might differ from the neoclassical norm. This possibility will be highlighted as the basic model is extended to see how bargaining influences the rate of labor-capital substitution and the firm's growth rate.

VI. A Long-Run Model

In order for quasi rent to accrue to the firm, there must be factors that are firm specific and not easily marketable. As Marshall underscores in his view of employer-employee connections, such rent-earning factors are frequently organizational in nature. In extending the basic model to a full-fledged long-run version, I explicitly put the firm's organizational capital in the production function, underscore the firm specificity of the labor pool, and let bargaining determine the capital-labor ratio as well as the long-run growth rate of the firm. In what follows, physical capital, simply called capital hereafter, will be denoted by K, and organizational capital by Z.

A natural specification of the firm's underlying long-run technology is a constant returns to scale in *all* inputs.[25] But, for a given amount of organizational capital, the firm's scale of operation will be delimited. For tractability, I restrict the functional form

[24] In the above medium-term interpretation, labor's objective may have to be modified to accommodate the possibility of labor pool reductions, because the dismissal from the labor pool introduces an additional risk. Efficient risk sharing here requires severance payments to the dismissed. It can then be shown that the efficient provision of severance payments entails $Esf'(n(s))=u^{-1}(V^*)$ if the labor pool is to be reduced. In other words, the worker's concern for severance ensures the average productive efficiency of the labor pool regardless of relative bargaining power. The short-run aspects of the bargained contract remain unaltered. The basic model applies straightforwardly to the case of a labor pool expansion.

[25] The notion of constant-returns technology has been proven particularly useful in the presence of firm-specific resources. For example, it is used by Edward Prescott and Michael Visscher (1980) to explain a competitive firm's rate of growth of its organizational capital; it is related to Robert Lucas' (1967) adjustment-cost model. In a model of general competitive economy, Lionel McKenzie (1959) argues that the constant returns is the most appropriate specification of the firm's production when rents accrue to factors that are private to the firm and not marketed.

of the long-run technology to

$$(7)\quad H(n,K,Z)=\min\{F(n,K),Z\}.$$

It is assumed that $F(\cdot)$ is linear homogeneous in (n,K) so that H is linear homogeneous in (n,K,Z). Organizational capital thus determines the firm's scale of operation. I also posit a Penrose hypothesis of the firm's growth cost as follows. It is assumed that the firm must allocate $T(g)Z$ in order to expand its organizational capital at rate g, that is, from Z to $(1+g)Z$. Also assume $T(g)$ to be strictly monotone and convex: $T'>0$ and $T''>0$. The growth cost includes the cost of the firm-specific training, information acquisition, marketing, and adjustment cost in input acquisition. Since the growth cost has to be paid out of the firm's revenue, the growth rate affects the present discounted value (PDV) of the remunerable rents. As such, g can be a bargaining instrument.

Since we are in a multiperiod framework, let $N(t)$ and $K(t)$ be the labor pool size and the capital input level that are chosen *ex ante* to the tth period. Short-run variables, which are contingent on the uncertainties within the period, are indexed by t and s. Let E_t be the expectation taken over s with respect to the tth period probability distribution of s. Assume that the firm has an access to a perfect capital market across t and s. Also assume a long-run stationary environment so that $E_t=E$ for all t; the rental price of capital input r and the workers reservation wage income k are once again constant for all t and s. For convenience, management's horizon is infinite, though the individual worker's is finite τ periods.

By the same insurance logic as before, we have $w(s,t)=c(s,t)+k$ for all s in t. Similarly, because the firm is risk neutral, $w(t)=w(s,t)$ and $c(t)=c(s,t)$. By using Baily's original logic, it can further be shown that an efficient contract requires income to be constant over time; thus, $w=c+k$, $w=w(s)$ and $c(s)=c$ for all t. Consequently, the worker's PDV of a utility stream can be written as $u(c+k)[D]$ where $[D]=\Sigma_{t=0}^{\tau}(1+\rho)^{-t}$, and ρ is the discount rate. The profit equation with an efficient contract can then be expressed as

$$(8)\quad \Pi(s,t)=sF(n(s,t),K(t))-kn(s,t)$$
$$-cN(t)-rK(t)-T(g(t))Z(t),$$

and the constraints are

$$(9)\quad n(s,t)\leqq N(t)\quad \text{for all } s \text{ in each } t,$$

$$(10)\quad F(n(s,t),K(t))\leqq Z(t)$$
$$\text{for all } s \text{ in each } t$$

$$(11)\quad (1+g(t))Z(t)=Z(t+1)\quad \text{for each } t,$$

given the initial condition $Z(0)$. An efficient contract can then be considered a solution to the long-run PMF program that maximizes $\Sigma_{t=0}^{\infty}E\Pi(s,t)/(1+\rho)^{t}$ subject to the above constraints and $u(c+k)[D]\geqq V$. Labor's utility requirement is expressed as above because of *ex ante* worker homogeneity and the long-run stationarity, and by implicitly assuming smoothly overlapping generations of recruitment and retirement.

By inspecting the first-order conditions on $n(s,t)$, $N(t)$, $K(t)$, $g(t)$, c, we recognize that $N(t)/K(t)$, $n(s,t)/K(t)$ and $g(t)$ are all t-invariant, independent of the initial condition $Z(0)$. Also, $Z(t)=F(N(t),K(t))$ and $Z(t)$ and $K(t)$ grow at the same rate. All of that follows from the linear homogeneity of F. We can therefore rewrite the efficient program by dropping t as follows. Define $L=N(t)/K(t)$, $l(s)=n(s,t)/K(t)$, and $f(l(s))=F(l(s),1)$. After substituting them into (8)–(11), we can characterize the efficient contract as $\{l(s),L,g,c,K(0)\}$ that solves

$$(\text{LP})\quad \text{Max } ((1+\rho)/(\rho-g))E[sf(l(s))$$
$$-kl(s)-cL-r-T(g)f(L)],$$

subject to $l(s)\leq L$ for all s, the utility constraint, and the initial condition $K(0)f(L)=Z(0)$.[26] Only $K(0)$ depends on $Z(0)$, and it is

[26] A finite maximum for the PDV of $E\Pi(s)$ requires $\rho>g$.

immaterial to the determination of $l(s)$, L, g, and c.[27]

Because the structure of (LP) is essentially the same as that of (P) in the basic model, (LP) inherits essential properties of Lemmas 1–4. A main proposition which follows from (LP) is that *$\partial g/\partial V<0$ and $\partial L/\partial V<0$.* Proposition 1 is now generalized to read, *strong labor obtains a higher remuneration and lower layoff probabilities in all states by bargaining for a higher capital-labor ratio and a slower growth rate*. Other forces identified in the basic model are also present although interactions between the labor-capital ratio and the growth rate sometimes make comparative statics complex. For example, we can readily confirm $[\partial L/\partial k]_V>0$, and also $[\partial g/\partial k]_V>0$. Similarly, an increase in *MPS*-uncertainty unambiguously accelerates the firm's growth. But, the effect of an *MPS* on L becomes ambiguous even for the *PMF* because it depends explicitly on the magnitudes of f''' and f''.

VII. Interpretation

I now provide some economic intuition to these results, since it may seem that, under efficient bargaining, rational bargainers would readily agree on a choice of (N,K,g) to maximize the whole of the expected business rent, only then to haggle over the division. To begin, however, the decision on (N,K) cannot be dichotomized from eventual claims to the rent. This is because the worker's gain depends on two things: the size of the rent, and the number of workers with whom he shares labor's claim on the rent. Further, there is a menu of capital-labor substitutions that can generate the same level of rent. Thus, for a given "iso-rent" curve, labor's per capita rental share increases as capital is used more intensively vis-à-vis the labor input. So, labor has a reason to prefer a low N/K. Profit-conscious management, on the other hand, will adopt a more capital-intensive method if the wage cost $(w=c+k)$ is high. Thus a stronger labor bargaining position results in a higher w and a lower N/K. Management of course wants a lower w, which entails a higher N/K ratio to maximize profits.[28]

Labor's preference for a slower growth rate can similarly be explained. With constant returns to scale in (N,K), the remuneration per worker depends only on the N/K ratio, not at all on the firm size. Yet, to have a larger firm size, a greater portion of the current revenue must be allocated to the growth cost, depressing labor's per capita remunerable share. Management, on the other hand, is interested in the *PDV* of the *total* profit stream. With the constant-returns technology, total profits can increase without bound unless the growth cost curve becomes increasingly steep. As a result, management is more growth oriented than labor.

Although I have used constant-returns technology to delineate the conflict of interest, the explanation is robust to changes in my assumptions about the firm's production technology.[29] Let me hasten to add that my argument must be modified if the assumption of either the competitive firm or the homogeneous labor is violated. If the firm is monopolistic in the output market, the per capita rental remuneration to labor depends on total revenue, hence, on the firm size. Likewise, if workers are hierarchically organized in seniority/promotion ladders, the worker's income stream depends on the rate and probability of promotion, hence on the firm's growth rate. Despite these qualifications, the fundamental conflict between labor

[27] Detailed derivations of the maximization conditions and proofs of various assertions in this section are available upon request.

[28] In anchoring the rationale for each party's N/K preference, I differ from the traditional explanation in two respects. Namely, (i) the labor-capital ratio itself is the direct object of bargaining, and (ii) the short-run employment of labor is independent of the bargained wage rate. This logic would have to be modified if the relative bargaining power were affected by the ratio N/K.

[29] If the underlying technology everywhere exhibits decreasing returns, labor will prefer to have the smallest possible N and g; that is, $N\to 1$ and $g\to 0$ as $R\to 0$. Consequently, a strictly concave production function accentuates the preceding logic of conflict. Only to the extent that increasing returns are not exhausted can collinearity of goals between labor and management occur. Finally, note that a nontrivial solution to internal bargaining exists only if a constant-returns-to-scale technology prevails at least locally in the solution.

and management remains regarding the determination of capital-labor inputs and the firm's growth rate as long as labor's maximand represents the per member share of the rent.[30] Once a choice of (N, K, g) is settled, it becomes in the bargainers' mutual interest to utilize the firm's resources in the most productive manner. Therefore, the marginal revenue product rule obtains in the short run.

VIII. Concluding Remarks

The view taken in this paper has been that the earnings potential of the firm depends on the firm-specific organizational capital, a composite of human and capital inputs bound together by long-term contracts. By assuming the bilateral efficiency between labor and management, I have explained how short-run commonality as well as long-run differences among firms results from the varied circumstances of internal bargaining.

Although the model is not specifically meant to study the impact of typical labor unions, the results have been generally consistent with empirical findings on firms with strong labor unions. According to Freeman and Medoff, labor-union studies tend to confirm that wages are less responsive to labor market conditions, there is much more cyclical adjustment via temporary layoffs, firms have lower labor-capital ratios, and firms have lower rates of profit per unit of capital, in the unionized sector than in the nonunionized sector. On the other hand, my model does not predict Gibrat's law, which hypothesizes that the firm's growth rate is independent of its size. The results will not bear out such a law unless various basic conditions offset the effect of relative bargaining power across firms. However, even if a firm's growth rate were not amenable to bargaining, being given exogenously for such a law, all essential aspects of my conclusions on the rest of the bargaining variables would still hold.

Throughout the discussion, it has been assumed that workers are homogeneous. In reality, however, firm growth and worker life cycle planning are related, and the issues of middle-run adjustment must be dealt with more seriously. Consideration of the worker's life cycle aspects may undercut the robustness of the results. The development of seniority and hierarchical wage structures becomes inevitable because more experienced workers come to possess a more substantive part of the firm's organizational capital. In short, vertical division of rent within labor becomes intertwined with the formation of organizational capital. In this regard, the results due to Aoki (1982) and Hiroyuki Odagiri (1980) are suggestive. My analysis is complementary to theirs by isolating the forces affecting the long-run horizontal rental division between labor and management. I have presented a formal, unified model of the firm under uncertainty by incorporating the labor-contractual aspects of bargaining into the Marshallian theory of the firm.

[30] See J. E. Meade (1972) for the convention of aggregating heterogeneous workers and calculating the per member share in terms of labor-efficiency units.

REFERENCES

Aoki, Masahiko, "A Model of the Firm as a Stockholder-Employee Cooperative Game," *American Economic Review*, September 1980, *70*, 600–10.

______, "Equilibrium Growth of the Hierarchical Firm: Shareholder-Employee Cooperative Game Approach," *American Economic Review*, December 1982, *72*, 1097–110.

Azariadis, Costas, "Implicit Contracts and Underemployment Equilibria," *Journal of Political Economy*, December 1975, *83*, 1183–202.

Baily, Martin Neil, "Wages and Employment Under Uncertain Demand," *Review of Economic Studies*, January 1974, *41*, 37–50.

Epstein, Larry, "Production Flexibility and the Behavior of the Competitive Firm Under Price Uncertainty," *Review of Economic Studies*, June 1978, *45*, 251–61.

Freeman, Richard B. and Medoff, James L., "The Impact of Collective Bargaining: Illusion or Reality?," in Jack Stieber et al., eds., *U.S. Industrial Relations 1950–1980: A Critical Assessment*, Madison: IRRA, University of Wisconsin, 1981, ch. 2, 47–97.

Grossman, Sanford J. and Hart, Oliver D., "Implicit Contracts, Moral Hazard and Unemployment," *American Economic Review Proceedings*, May 1981, *71*, 301–07.

Hall, Robert E. and Lilien, David M., "Efficient Wage Bargains Under Uncertain Supply and Demand," *American Economic Review*, December 1979, *69*, 868–79.

Holmstrom, Bengt, "Equilibrium Long-Term Labor Contracts," *Quarterly Journal of Economics*, Suppl., 1983, *98*, 23–54.

Kalai, E., "Nonsymmetric Nash Solutions and Replications of 2-Person Bargaining," *International Journal of Game Theory*, 1977, *6*, 129–33.

Lucas, Robert E., Jr., "Adjustment Costs and the Theory of Supply," *Journal of Political Economy*, August 1967, *75*, 321–34.

McDonald, Ian M. and Solow, Robert M., "Wage Bargaining and Employment," *American Economic Review*, December 1981, *71*, 896–908.

McKenzie, Lionel W., "On the Existence of General Equilibrium for a Competitive Market," *Econometrica*, January 1959, *27*, 54–71.

Marshall, Alfred, *Principles of Economics*, 8th ed., London: Macmillan (Student Edition), 1969.

Meade, J. E., "The Theory of Labour-Managed Firms and of Profit-Sharing," *Economic Journal*, March 1972, *82*, 402–28.

Miyazaki, Hajime, "A Marshallian Theory of the Firm," Research Paper No. 28, Workshop on Factor Markets, Stanford University, September 1982.

______ **and Neary, Hugh M.,** "The Illyrian Firm Revisited," *Bell Journal of Economics*, Spring 1983, *14*, 259–70.

Odagiri, Hiroyuki, *The Theory of Growth in a Corporate Economy: An Inquiry into Management Preference, R&D, and Economic Growth*, Cambridge: Cambridge University Press, 1980.

Oi, Walter Y., "The Desirability of Price Instability Under Perfect Competition," *Econometrica*, January 1961, *29*, 58–64.

Prescott, Edward C. and Visscher, Michael, "The Organization Capital," *Journal of Political Economy*, June 1980, *88*, 446–61.

Riddell, W. Craig, "Bargaining Under Uncertainty," *American Economic Review*, September 1981, *71*, 579–90.

Rosen, Sherwin, "Unemployment and Insurance," Working Paper No. 1095, National Bureau of Economic Research, March 1983.

Svejnar, Jan, "On the Theory of a Participatory Firm," *Journal of Economic Theory*, August 1982, *27*, 313–30.

Vanek Jaroslav, *The General Theory of Labor-Managed Market Economies*, Ithaca: Cornell University Press, 1970.

Ward, Benjamin, "The Firm in Illyria: Market Syndicalism," *American Economic Review*, September 1958, *48*, 566–89.

Innovation, Market Structure, and Welfare

By PANKAJ TANDON*

Many writers subscribe to Joseph Schumpeter's view that, while perfectly competitive firms allocate resources efficiently in a static sense, they perform poorly when it comes to innovation.[1] From this point of view, the optimal form of market structure is unlikely to be perfect competition, but some other type of dynamic competition which includes significant elements of monopoly. Recently, considerable effort has been focused on modelling Schumpeter's notion of competition. Perhaps best exemplified by the 1980 work of Partha Dasgupta and Joseph Stiglitz (hereafter D-S),[2] this approach views free entry to the *R&D* game, rather than to production, as the relevant notion of dynamic competition.[3] Thus the market structure in production activities is endogenous in the model.

This paper extends the D-S approach to discuss the tradeoff between static and dynamic efficiency. The question asked is: what is the optimal market structure or optimal degree of concentration? The purpose is to compare the modified notion of competition (what D-S call a "free-entry oligopoly") with other types of blocked-entry oligopoly. A different way of stating the question then is: *are barriers to entry, in addition to those created by R&D, desirable*?

I show that the D-S approach will answer this question in the affirmative. However, a careful examination of the tradeoffs reveals a rather stronger result. It can be argued that, at the level of entry characteristic of the free-entry oligopoly, there may be no tradeoff at all! Purely static considerations are shown to lead a social welfare maximizer to argue for *increased* concentration. The fundamental reason is still the tendency towards scale economies that *R&D* results create.[4] By entering the industry, the marginal firm will inhibit the reaping of these scale benefits by inframarginal firms. Thus the marginal firm will in general make a net negative contribution to social welfare, even when we disregard the further dynamic effect on *R&D* incentives. Of course, once the industry is more concentrated than the free-entry outcome, this "perverse" static effect may begin to disappear. The result is similar in some respects to that of Stiglitz (1981) for the case of potential competition, although the driving force is different.

To demonstrate the point, an illustrative model of the D-S type is developed. It is found that the free-entry outcome performs relatively worse for industries that are characterized by high levels of technological opportunity.[5] Some simple numerical calcu-

*Assistant Professor of Economics, Boston University, 270 Bay State Road, Boston, MA 02215. The idea for this paper occurred to me at a seminar by Michael Spence, to whom I am grateful. I have received helpful comments from members of the B.U. micro theory research workshop, especially Randy Ellis, Michael Manove, and Ingo Vogelsang, and from an anonymous referee.

[1] The "Schumpeterian tradeoff" is referred to repeatedly. See, for example, Paolo Sylos-Labini (1969), F. M. Scherer (1980), C. C. von Weizsacker (1980), Richard Nelson and Sidney Winter (1982), and Morton Kamien and Nancy Schwartz (1982).

[2] There is a large and growing literature in this area. See, for example, Kamien and Schwartz (1972, 1976), Glenn Loury (1979), Carl Futia (1980), and my 1983 article. For a first attempt at empirical work in this area, see Richard Levin (1981). An alternative approach involves the use of simulation models; see Nelson and Winter (1977).

[3] Readers will recognize a similarity of this approach with the notion of contestable markets discussed most recently by William Baumol (1982) and by Baumol, John Panzar, and Robert Willig (1982). For an interesting comparison of the Schumpeterian with the Marxian notion of competition, see John Elliott (1980).

[4] This was noted, for example, by Robert Wilson (1975).

[5] In such industries, the long-term gains from dynamically efficient innovation become of paramount importance; consequently, the optimal market structure would consist of a small number of firms. The driving force behind such a result is what Scherer (1972) has called the "Lebensraum effect." Firms performing *R&D* must at least break even. They derive their profits from

lations carried out in Section II correspondingly suggest that the optimal degree of concentration rises as technological opportunities improve. It is found, however, that except for very high values of the technological opportunity parameter, the optimal degree of concentration will typically involve more than one firm. Thus although the free-entry outcome results in *excessive* duplication, it is seen that not *all* duplication is entirely wasteful. The reason is that the duplication of research in this second best world can indicate reduced static deadweight losses in the industry.[6]

I. A Simple Illustrative Model

The model used here is a modified version of the one proposed by Dasgupta and Stiglitz.[7] Consider the market for some well-defined new product. Consider the linear approximation to the market demand for this product:

$$p(Q)=a-bQ.$$

Let the average cost of production be c, a function of the amount of research, x, done on the product, but independent of output Q. For convenience, call the difference $(a-c)$ the "cost reduction" and represent it by B. Since c is a function of x, and a is a constant, B is also a function of x. Assume that this function is of a constant elasticity form,[8] so that

$$(1)\qquad B(x)=\beta x^{\alpha}.$$

The strategy of this section will be to look at the equilibrium outcomes under free entry and blocked entry, and to compare them to the socially optimal outcomes. In this manner, the value of a welfare index can be compared for different levels of entry and the optimal number of firms computed.

A. *The Social Problem*

Suppose society wishes to maximize net social benefit. The usual notion of consumer's surplus will be used to measure consumption benefits. If society can produce this product at a cost c, the optimum price to charge for it would be c. The quantity demanded at this price would be B/b. The consumer's surplus generated by reducing the cost of production to c would then be $B^2/2b$. If this represents a per annum benefit that extends indefinitely into the future, we may aggregate these future returns using the social discount rate, r. In this case, net social welfare may be written

$$(2)\qquad W=(B^2/2rb)-x.$$

The social problem then is to choose x in order to maximize W. Note that, since there is no uncertainty in this model, it is optimal

the market power with which research results endow them. In industries where it is socially desirable that large amounts of research be done— because research is highly productive—firms must be allowed a commensurate degree of market power. In industries where *R&D* is not particularly effective, more attention can be paid to static efficiency, and the free-entry outcome is comparatively better.

[6] Note that this defense of duplicative research is different from the argument I advanced in my 1983 paper. The argument there was that duplication is not always wasteful since, in the presence of uncertainty, it is a reasonable way to raise the probability of success in research.

[7] The main difference is that, whereas D-S assume the demand curve to have constant elasticity, the demand curve is taken here to be linear. The two assumptions seem in some sense to be equally arbitrary. On theoretical grounds, the constant elasticity assumption is superior, since it yields a consistent welfare index. On practical grounds, however, it is impossible to use. When the elasticity of demand is less than one, the index is negative. This obviously makes it impossible to make cardinal welfare comparisons of the kind attempted here. Further, it is well known that, if the elasticity of demand is less than unity, no monopoly equilibrium will exist. This would eliminate the possibility of a complete set of comparisons. The present framework has been used to analyze optimal patents (see my 1982a article) and has the virtue of permitting all the necessary comparisons. It should be noted, however, that the use of consumer's surplus as a welfare index has not been shown to be appropriate in general, so that the results below must be interpreted with great caution.

[8] Note that this formulation is different from that of D-S, who consider the case where $c(x)=\beta x^{-\alpha}$. Again, the two assumptions seem equally arbitrary and differ only by a constant term. I make mine for the convenience of computations presented in the next section.

for society to have only one firm. This is the sense in which this is a natural monopoly problem.

Using the functional form (1) for $B(x)$, it can be shown that the solution to the social problem involves

$$(3)\quad x_s=\left(\alpha\beta^2/rb\right)^{1/(1-2\alpha)},$$

$$(4)\quad B_s=\beta\left(\alpha\beta^2/rb\right)^{\alpha(1-2\alpha)},$$

$$(5)\quad Q_s=(\beta/b)\left(\alpha\beta^2/rb\right)^{\alpha/(1-2\alpha)},$$

$$(6)\quad W_s=((1-2\alpha)/2\alpha)\left(\alpha\beta^2/rb\right)^{1/(1-2\alpha)}.$$

This set of results is consistent with those of Dasgupta and Stiglitz. Given a value of the demand choke point, a, the demand slope parameter may be interpreted as indicating the size of the market, with a higher absolute slope indicating a smaller market. Equations (3)–(6) show that the optimal levels of *R&D* and output both rise as the market expands (b falls). Further, more productive research (characterized by higher β) also calls for higher *R&D* and output. This is all in accord with intuition.

B. *Free-Entry Oligopoly*

Following Dasgupta-Stiglitz and other recent work in this area, the competitive situation will be taken to be one where there is free entry, but not necessarily a large number of infinitesimal firms. Because *R&D* is something like a fixed cost of entry, firms must earn quasi rents in equilibrium that cover this cost. Thus in the product market there will be an oligopoly. This approach is similar to the familiar monopolistic competition model, and has been used recently in several different contexts (see, for example, Steven Salop, 1979).

The usual type of Cournot-Nash outcome will be treated as the equilibrium concept. Thus any firm that enters will choose its *R&D* spending and its output level to maximize profits, assuming other firms will not alter their behavior. Entry occurs as long as there are positive profits. The net present value of the ith firms' profits will be given by

$$(7)\quad \pi_i=(1/r)\left[a-b\left(Q_i+\sum_{j\neq i}Q_j\right)-a+B(x_i)\right]Q_i-x_i.$$

Concentrating on symmetric equilibria,[9] the first-order conditions for (7) to reach a maximum are

$$(8)\qquad (1/r)Q_iB'(x_i)=1,$$

$$(9)\qquad (n+1)bQ_i=B(x_i),$$

where n is the number of firms that have entered. A third condition that characterizes the market equilibrium is that no further entry is profitable. For the sake of analytical convenience, take this to imply $\pi_i=0$ for all i. This ignores the integer problem, but really does not cause the model to become unrealistic. Thus

$$(10)\quad (1/r)\left(B(x_i)-nbQ_i\right)Q_i=x_i.$$

Combining (8), (9), and (10) yields the interesting result that the equilibrium number of firms n^* will be given by

$$(11)\qquad n^*=(1-\alpha)/\alpha,$$

where $\alpha=(dB/B)/(dx_i/x_i)$, the elasticity of the cost reduction function. Note that the result (11) does not depend on any particular functional form for $B(x_i)$.

We may interpret α as a measure of technological opportunity. High α indicates that research is productive in producing cost reductions. Condition (11) says that the number of firms that will enter in competitive equilibrium is inversely related to technological opportunity. This is in line with the re-

[9] In general, the equilibrium may not be symmetric. An asymmetric outcome is obtained, for example, by M. Therese Flaherty (1980). Dasgupta and Stiglitz did assume a symmetric outcome. For the basic argument here, symmetry is not necessary but only a desirable simplification.

sults of D-S, although still somewhat surprising. One might expect that high possible payoffs to *R&D* might tempt many firms to enter. The factor dominant in this model, however, is that *R&D* serves as a barrier to entry. We will see in (18) below that, in industries marked by high levels of technological opportunity, *R&D* per firm will be high; thus such industries are characterized by higher entry barriers, and greater concentration.

Two other points about (11) deserve mention. First, and somewhat surprisingly, note that n^* does not depend upon the discount rate. Interestingly, this result persists in the constant elasticity demand case. A reduction in the discount rate might be expected to encourage entry since the value of a fixed revenue stream would rise. However, the cost of entry also would rise since firms would increase their *R&D* spending. Second, note that (11) does not contain any demand parameter. This result is akin to the D-S finding that the number of entrants did not depend on the size of the market, but is not robust with respect to different cost or demand conditions.[10]

To complete the discussion of this section, the outcome of equations (8)–(10), assuming the functional form (1), should be presented. In free-entry equilibrium, the level of *R&D* spending per firm, the "cost reduction" and the industry output level are computed to be

$$(12)\quad x^* = (\alpha^2\beta^2/rb)^{1/(1-2\alpha)},$$

$$(13)\quad B^* = \beta(\alpha^2\beta^2/rb)^{\alpha/(1-2\alpha)},$$

$$(14)\quad Q^* = ((1-\alpha)\beta/b)(\alpha^2\beta^2/rb)^{\alpha/(1-2\alpha)}.$$

From equation (2), recall that the present value of the potential gross social welfare gain is given by $B^2/2rb$. From this must be subtracted the triangle of deadweight loss due to the static inefficiency of price exceeding marginal cost, and of course the cost of research. Now in the linear demand case, a Cournot oligopoly produces $n/(n+1)$ times the competitive output. Thus

$$Q^* = n/(n+1)(B/b).$$

The welfare index in this case is given by

$$(15)\quad W = B^2/2rb - B^2/2rb(n+1)^2 - nx.$$

Using equations (11)–(14) and simplifying yields

$$(16)\quad W^* = (1/2)((1-\alpha)/\alpha)^2(\alpha^2\beta^2/rb)^{1/(1-2\alpha)}.$$

A comparison between these results and the social optimum, represented by equations (3)–(6) will be shown in the next section. However, a couple of observations are in order at this point. First, consider the relationship between *R&D* spending per firm and market structure. As argued earlier, we should observe high *R&D* per firm associated with concentration. Differentiating (11) we have

$$(17)\quad dn^*/d\alpha = -1/\alpha^2 < 0.$$

Increased concentration is associated with high α, that is, high technological opportunity. From (12) we have

$$(18)\quad dx^*/d\alpha = (2x^*/(1-2\alpha)) \times (1/\alpha + \ln x^*) > 0,$$

assuming $\alpha < 1/2$, which is certainly a reasonable assumption.[11] Thus high *R&D* spending per firm is associated with high technological opportunity. Combining (17) and (18) we get the result that high *R&D* per

[10] Although the parameter b cannot truly be interpreted as a "market size" parameter, it is curious that n^* does not depend on it. The parameter a in the inverse-demand curve may be thought of as indicating size. However, the model here implicitly normalizes with respect to a, rendering it difficult to analyze the effects of changing size on n^*.

[11] See Zvi Griliches (1973) for a discussion of the problems associated with estimating α and of some of the estimates that have been made. A reasonable estimate is 0.1; the highest estimate has been 0.12 by Edwin Mansfield (1965).

firm is associated with greater concentration. An obvious collary is that greater cost reductions are associated with greater concentration, which was Schumpeter's basic point, and matches the result of Dasgupta and Stiglitz. Also parallel to D-S, note from (3) and (12) that $x^* < x_s$; the free-entry oligopoly will always result in cost reductions that are too small relative to the socially managed industry.

Second, what about total *R&D* spending? Define $X^* = n^* x^*$. It is not immediately obvious how this varies with α, since $dn^*/d\alpha$ and $dx^*/d\alpha$ have opposite signs. However, it can be shown that $dX^*/d\alpha$ is positive.[12] Thus total *R&D* spending also rises with technological opportunity and concentration. If we compare X^* with x_s, we find from (3), (11), and (12) that $X^* < x_s$ always. Here, unlike the D-S model, even aggregate *R&D* spending remains smaller than the socially optimal level for one firm.

C. *Blocked-Entry Oligopoly*

A Cournot-Nash equilibrium for an industry with blocked entry is now examined. Suppose the fixed number of firms in the industry is n. Then (8) and (9) will be the typical firm's first-order conditions if we concentrate on symmetric equilibria. Solving these equations yields the equilibrium levels of *R&D* per firm, cost reduction, and industry output.

$$(19)\quad x_b = \left(\alpha\beta^2/rb(n+1)\right)^{1/(1-2\alpha)},$$

$$(20)\quad B_b = \beta\left(\alpha\beta^2/rb(n+1)\right)^{\alpha/(1-2\alpha)},$$

$$(21)\quad Q_b = (n\beta/(n+1)b) \times \left(\alpha\beta^2/rb(n+1)\right)^{\alpha/(1-2\alpha)}.$$

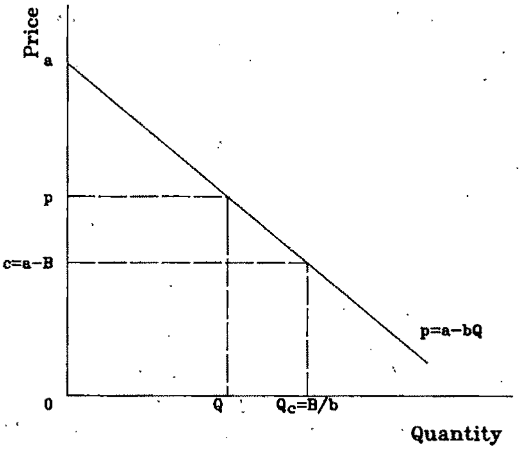


FIGURE 1

It is possible to show that, if n were given by (11) (i.e., were the free-entry number of firms), (19)–(21) would reduce to (12)–(14).

To construct the welfare index for this case, it is useful to examine Figure 1. In equilibrium, the cost of production is $c = a - B(x)$, and output is Q. Recall that in the linear demand case $Q = nQ_c/(n+1)$, where Q_c, the competitive output, is given by $Q_c = B/b$. This can be rearranged to yield $Q_c - Q = Q_c/(n+1)$. The welfare index can then be written as

$$(22)\quad W_b = B^2/2rb - (p-c)Q_c/2r(n+1) - nx_b.$$

It is now possible to take up the question of whether the free-entry outcome could be improved upon by greater concentration, solely on grounds of static efficiency. To examine the static efficiency effect[13] of there

[12] Using (17) and (18) we may write

$$dX^*/d\alpha = -x^*/\alpha^2 + ((1-\alpha)/\alpha) \times[(2x^*/(1-2\alpha))(1/\alpha + \ln x^*)].$$

This may be simplified to

$$dX^*/d\alpha = -(x^*/\alpha^2(1-2\alpha))[1+2\alpha(1-\alpha)\ln x^*]$$

which is positive for $\alpha < 1/2$.

[13] Note that I have not explicitly analyzed here the dynamic effect. It is clear that the remaining firms will have increased incentives to perform *R&D*. The cost of production will decline, price will fall, and industry output will increase. On the face of it, this constitutes a welfare improvement. In general, I believe it will be so. However, I wish to offer an interesting alternative possibility. It *may* be that the dynamic effect could actually work in the other direction, i.e., that it could lead to a decline in welfare relative to the static $(n-1)$-firm equilibrium. The reason is the traditional rent-seeking or common pool argument. In the $(n-1)$-firm static equilibrium, each firm will be making some extra-normal profit above its *R&D* spending. This profit of course

being one less firm at the free-entry level of concentration, we must rewrite W_b for $(n-1)$ firms, holding x_b (the *R&D* spending level per firm) and hence Q_c, c, and B constant. Keeping in mind that price will be higher by $(p-c)/n$, we find the change in welfare is

$$(23)\quad \Delta W_b = x_b(1+2n-2n^2)/2n^2.$$

Now note that for $n \geq 2$ the right-hand side of (23) is always negative. Thus we may state the basic proposition:

PROPOSITION 1: *At the free-entry level of concentration, static efficiency improves with concentration.*[14]

For purposes of further comparison, it is useful at this point to rewrite (22) in terms of the parameters of the model. Using (19)–(21), we obtain

$$(24)\quad W_b = [n(n+2)/2\alpha(n+1)-n] \times (\alpha\beta^2/rb(n+1))^{1/(1-2\alpha)}.$$

The optimal degree of concentration, taking into account static and dynamic effects, would be the value of n that maximizes (24). This is not easy to do analytically, but can be done numerically as the following section shows.

II. Welfare Comparisons

There are three sets of equations derived above. Equations (3)–(6) represent the outcomes of social management; they will be taken as the ideal values of the different variables. Equations (12)–(14) and (16) represent the outcomes in a free-entry oligopoly, which is the case nearest to a competitive equilibrium that has been discussed. Finally, equations (19), (21), and (24) represent the outcomes in an industry with blocked entry.

This section makes comparisons between the socially optimum outcomes on the one hand and the free-entry and blocked-entry outcomes on the other. Comparisons are also possible between the blocked-entry outcomes for different values of n, so that it is possible to speak then of the "optimal" market structure.

Dividing the respective free-entry outcomes by equations (3)–(6) yields the following ratios:

$$(25)\quad x^*/x_s = \alpha^{1/(1-2\alpha)}$$

$$(26)\quad B^*/B_s = \alpha^{\alpha/(1-2\alpha)}$$

$$(27)\quad Q^*/Q_s = (1-\alpha)\alpha^{\alpha/(1-2\alpha)}$$

$$(28)\quad W^*/W_s = (1-a)^2\alpha^{1/(1-2\alpha)}/\alpha(1-2\alpha).$$

We see that the relative performance of the free-entry outcomes to the social optima is determined entirely by the technological opportunity parameter α. Table 1 lists the values of these ratios for different values of α, which is taken to range in value from 0.2 to 0.01. A value of 0.2 would indicate a dynamic industry characterized by high technological opportunity. It is seen from the table that, although the level of *R&D* spending as a proportion of the ideal declines as α falls, the welfare index rises. There are two significant reasons for this. First the size of the output distortion falls as the number of firms increases, that is, as α falls. This effect is clearly captured in the fifth column of Table 1, which shows the free-entry output level getting closer to the ideal level as α falls. In principle, the effect ought to be counteracted by the increased dynamic efficiency losses. The second point, however, is

belongs in our measure of social welfare. It is possible that this profit will entirely be dissipated in rivalry over relatively unproductive research lines as has been argued by Yoram Barzel (1968), Jack Hirshleifer (1971), and others. In my 1983 article, I constructed a specific example to demonstrate this possibility in a model where the results of research are uncertain. I do believe, however, that in general the dynamic effect will argue for increased concentration.

[14] The sharpness of the result here does depend a bit on the restrictive functional forms assumed. However, I believe the basic point is quite general. In my 1982b article, I argued the point more generally and also considered the case of constant elasticity demand. In this case, the result continues to hold unambiguously even when the free-entry equilibrium sustains only three firms. The result also has parallels in the literature on spatial competition. See Curtis Eaton and Richard Lipsey (1978, 1979), Scherer (1979), and Salop.

TABLE 1—COMPARISON OF *R&D*, COST REDUCTION, OUTPUT, AND WELFARE UNDER FREE-ENTRY OLIGOPOLY WITH THE SOCIAL OPTIMUM

n^*	α	$\frac{x^*}{x_s}$	$\frac{B^*}{B_s}$	$\frac{Q^*}{Q_s}$	$\frac{W^*}{W_s}$
4	.200	.068	.585	.468	.365
6	.143	.066	.678	.581	.472
9	.100	.056	.750	.675	.570
12	.077	.048	.792	.731	.632
15	.063	.042	.820	.769	.676
20	.048	.035	.852	.811	.728
25	.038	.029	.873	.839	.763
30	.032	.025	.888	.860	.790
40	.024	.020	.909	.887	.827
50	.120	.017	.923	.905	.852
100	.010	.009	.954	.945	.911

Note: n^* is the number of firms that enter under free entry.

that, although the *R&D* ratio does decline with α, the size of the cost reduction as a proportion of the ideal *rises* as α falls. This is entirely plausible with a concave cost-reduction function.

It has already been shown that the number of entering firms varies inversely with α. High opportunity industries are characterized by concentration. From (16) it is possible to show that the welfare index W^* varies directly with α—this would indicate that industries characterized by greater concentration also have associated with them a higher value of the welfare index. However, this should not be taken as an argument in favor of concentration. The driving force behind these results is the technological opportunity parameter α. Industries characterized by high technological opportunities create more social welfare simply because technological opportunities are high. What Table 1 shows clearly is that, *relative to the ideal welfare level* W_s, the industries characterized by *less* concentration perform considerably better than those that are more highly concentrated, that is, less concentrated industries come closer to realizing the maximum possible social gains than do more concentrated industries. This is because they produce a closer-to-ideal output (fifth column, Table 1) and also generate cost reductions closer to the ideal (fourth column). It must be kept in mind that this discussion has applied only to the free-entry case.

Let us turn now to the case of blocked entry, in order that we may find the optimal level of concentration. Using equations (6) and (24), it is possible to construct for our special case the ratio of the welfare index under blocked entry to the ideal maximum:

$$(29)\quad W_b/W_s=\big[(n(n+2)-2\alpha n(n+1))\\ /(1-2\alpha)\big](1/(n+1))^{(2-2\alpha)/(1-2\alpha)}.$$

This is a complex equation that is not easy to interpret. It shows the welfare index ratio as a function only of α, the technological opportunity parameter, and n.

Table 2 shows the value of this welfare index ratio for different values of α and n. The first column shows the corresponding value of n^*, the number of firms that would enter under free entry. For any given value of α, the welfare index is single peaked across the number of firms, with the peak occurring at smaller values of n for higher values of α. The peak of course represents the optimal number of firms. It is seen that this optimal number tends to be rather low and is always less than n^*. Further, it falls as the industry becomes more dynamic (α rises). This is consistent with the traditional view of dynamic vs. static efficiency. The greater the technological opportunities in an industry, the greater is the social payoff to the increased *R&D* incentives generated by concentration.

TABLE 2—VALUES OF THE WELFARE INDEX W_b/W_s FOR DIFFERENT VALUES OF α AND n

		Number of Firms											
n^*	α	1	2	3	4	5	6	9	15	20	30	50	100
4	.200	.577	.498	.422	.365	.322	.290	.226	.163	.135	.103	.074	.046
6	.143	.644	.630	.581	.538	.502	.472	.408	.336	.301	.256	.209	.158
9	.100	.682	.718	.696	.669	.643	.621	.569	.505	.471	.427	.376	.316
12	.077	.701	.761	.755	.739	.720	.703	.662	.608	.578	.538	.491	.433
15	.063	.712	.787	.791	.781	.768	.753	.722	.676	.650	.614	.572	.518
20	.048	.722	.813	.828	.825	.817	.807	.784	.749	.728	.698	.662	.616
25	.038	.728	.828	.849	.851	.847	.843	.823	.794	.777	.752	.721	.681
30	.032	.731	.839	.864	.869	.868	.865	.850	.826	.811	.790	.763	.728
40	.024	.736	.851	.882	.892	.893	.892	.884	.867	.855	.839	.818	.790
50	.020	.739	.859	.893	.905	.909	.909	.905	.892	.883	.869	.852	.829
100	.010	.745	.873	.915	.932	.940	.944	.947	.943	.939	.933	.924	.911

Note: See Table 1. The underlined values point out the optimal number of firms, n_b^*.

TABLE 3—COMPARISON OF THE FREE-ENTRY AND OPTIMAL OUTCOMES

α (1)	n^* (2)	n_b^* (3)	W^*/W_s (4)	W_b^*/W_s (5)	W^*/W_b (6)	x^*/x_b^* (7)	$\frac{n^*x^*}{n_b^*x_b^*}$ (8)
.200	4	1	.365	.577	.633	.216	.864
.143	6	1	.472	.644	.733	.174	1.044
.100	9	2	.570	.718	.794	.221	.995
.077	12	2	.632	.761	.830	.176	1.056
.063	15	3	.676	791	.855	.205	1.025
.048	20	3	.728	.828	.879	.162	1.080
.038	25	4	.763	.851	.897	.166	1.038
.032	30	4	.790	.869	.909	.140	1.050
.024	40	5	.827	.893	.926	132	1.056
.020	50	6	.852	.909	.937	.129	1.075
.010	100	9	.911	.947	.962	.095	1.056

Another interesting point to note in Table 2 is that the drop in the welfare index moving from the peak to lower concentration is more dramatic at high levels of α than at low levels. Now it may be that society wishes relatively low levels of concentration for reasons other than those modeled in this paper. The point to be made here is that the cost of such antitrust action will be greater in dynamic industries characterized by high technological opportunity than in relatively "static" industries. But curiously, these are precisely the industries where concentration will tend to be more pronounced, as was seen in Table 1. Thus an interesting paradox is created for antitrust legislators.

The peaks in the welfare index for different values of α are underlined in Table 2. The number of firms at these peaks is the optimal number of firms. Of course this optimal number is a function of α. In Table 3, the optimal outcomes are compared with the free-entry outcomes. For example, n^* denotes the number of firms in free entry and n_b^* denotes the optimal number under blocked entry. For comparative purposes, the proportion of the ideal maximum welfare captured in the respective cases is presented in columns 4 and 5. This leads to column 6, which shows the proportion of the blocked-entry optimal welfare captured under free entry. It is seen that free entry does relatively better for low values of α. The desirability of additional barriers to entry is greater in dynamic industries. What drives this result is that restricting entry will lead to higher *R&D* spending per firm. Column 7 shows the *R&D* spending per firm in free entry as a proportion of the spending under the optimal outcome. I do not attach any particular significance to the specific numbers here; they are sufficient to indicate the basic effect of concentration on *R&D* incentives. Last, column 8 shows the ratio of total *R&D* spending in the two cases. Curiously, this ratio stays fairly close to unity. Again, this may not have much significance or importance. It does remind us of one particular assumption made in the model, namely that the *R&D* of different firms is purely duplicative. This is of course not realistic and needs to be modified in subsequent work.[15]

III. Concluding Remarks

This paper has examined the Schumpeterian tradeoff using a simple framework and the familiar technique of calculating consumer's surplus. In a sense, this is an extension of the approach of Oliver Williamson (1968) who pointed out that economies of scale could be used as a defense of monopoly and suggested a similar approach to its measurement.[16] I have shown here that free entry to the *R&D* game would lead to excessive entry, in the sense that an industry with fewer firms would be socially preferable. This was true even when free entry led to the entry of only a small number of firms. A simple model with specific functional forms indicated that the "optimal" market structure would in general involve few firms, particularly in industries characterized by high levels of technological opportunity. These results are also consistent with the Schumpeterian notion of competition.

The model used here is highly simplified, of course, and the conclusions accordingly limited. It was primarily my aim to examine the implications of the influential Dasgupta-Stiglitz approach to this problem. One important point that emerges is that it is not realistic to compare competition with monopoly—to use the usual characterization of the problem—in a model where all firms do the same research. Concentrated industries come out looking good in this paper since further entry adds no new knowledge. Of course, one of the key advantages of a more competitive environment is precisely that a greater diversity of ideas is allowed to flourish. Modelling this phenomenon is a key area of research in this field.

Let me note one other shortcoming of the model. The model has been essentially static, in that the technological opportunities are a one-shot deal. In fact, technological conditions in an industry are constantly changing. The model may be able to say something when the changes are exogenous. In this case, we might expect the industry structure to become less concentrated over time as technological opportunities are "used up." However, many of the changes may be endogenous, and intimately connected with market structure. The pyramiding of inventions is an important phenomenon that has received inadequate attention. Further research in this area is also of some importance in a proper understanding of the tradeoffs between competition and monopoly.

[15] For a model where firms do discover different things when doing the same amount of *R&D*, see my 1983 article.

[16] See also his update of the argument, Williamson (1977).

REFERENCES

Barzel, Yoram, "Optimal Timing of Innovations," *Review of Economics and Statistics*, August 1968, *50*, 348–55.

Baumol, William J., "Contestable Markets: An Uprising in the Theory of Industry Structure," *American Economic Review*, March 1982, *72*, 1–15.

______, **Panzar, John C., and Willig, Robert D.,** *Contestable Markets and the Theory of Industry Structure*, San Diego: Harcourt Brace Jovanovich, 1982.

Dasgupta, Partha S. and Stiglitz, Joseph E., "Industrial Structure and the Nature of Innovative Activity," *Economic Journal*, June 1980, *90*, 266–93.

Eaton, B. Curtis and Lipsey, Richard G., "Freedom of Entry and the Existence of Pure Profit," *Economic Journal*, September 1978, *88*, 455–69.

______ and ______, "The Theory of Market Preemption: The Persistence of Excess Capacity and Monopoly in Growing Spatial Markets," *Economica*, May 1979, *46*, 149–58.

Elliott, John E., "Marx and Schumpeter on Capitalism's Creative Destruction: A Comparative Restatement," *Quarterly Journal of Economics*, August 1980, *95*, 45–68.

Flaherty, M. Therese, "Industry Structure and Cost-Reducing Innovation," *Econometrica*, July 1980, *48*, 1187–211.

Futia, Carl, "Schumpeterian Competition," *Quarterly Journal of Economics*, June 1980, *94*, 675–95.

Griliches, Zvi, "Research Expenditures and Growth Accounting," in B. R. Williams, ed., *Science and Technology in Economic Growth*, New York: Wiley & Sons, 1973.

Hirshleifer, Jack, "The Private and Social Value of Information and the Reward to Inventive Activity," *American Economic Review*, September 1971, *61*, 561–74.

Kamien, Morton I. and Schwartz, Nancy L., "Timing of Innovations under Rivalry," *Econometrica*, January 1972, *40*, 43–60.

______ and ______, "On the Degree of Rivalry for Maximum Innovative Activity," *Quarterly Journal of Economics*, May 1976, *90*, 245–60.

______ and ______, *Market Structure and Innovation*, Cambridge: Cambridge University Press, 1982.

Levin, Richard C., "Toward an Empirical Model of Schumpeterian Competition," Working Paper No. 43, Series A, Yale University School of Organization and Management, July 1981.

Loury, Glenn C., "Market Structure and Innovation," *Quarterly Journal of Economics*, August 1979, *93*, 395–410.

Mansfield, Edwin, "Rates of Return from Industrial Research and Development," *American Economic Review Proceedings*, May 1965, *55*, 310–22.

Nelson, Richard R. and Winter, Sidney G., "Dynamic Competition and Technical Progress," in Bela A. Balassa and Richard R. Nelson, eds., *Economic Progress, Private Values and Public Policy: Essays in Honor of William Fellner*, Amsterdam: North-Holland, 1977.

______ and ______, "The Schumpeterian Tradeoff Revisited," *American Economic Review*, March 1982, *72*, 114–32.

Salop, Steven C., "Monopolistic Competitition with Outside Goods," *Bell Journal of Economics*, Spring 1979, *10*, 141–56.

Scherer, F. M., "Nordhaus' Theory of Optimal Patent Life: A Geometric Reinterpretation," *American Economic Review*, June 1972, *62*, 422–27.

______, "The Welfare Economics of Product Variety: An Application to the Ready-to-Eat Cereals Industry," *Journal of Industrial Economics*, December 1979, *28*, 113–34.

______, *Industrial Market Structure and Economic Performance*, 2d ed., Chicago: Rand McNally, 1980.

Spence, A. Michael, "Cost Reduction, Competition and Industry Performance," Discussion Paper Number 897, Harvard Institute of Economic Research, April 1982.

Stiglitz, Joseph E., "Potential Competition may Reduce Welfare," *American Economic Review Proceedings*, May 1981, *71*, 184–89.

Sylos-Labini, Paolo, *Oligopoly and Technical Progress*, Cambridge: Harvard University Press, 1969.

Tandon, Pankaj, (1982a) "Optimal Patents with Compulsory Licensing," *Journal of Political Economy*, June 1982, *90*, 470–86.

______, (1982b) "Innovation, Market Structure and Welfare," Department of Economics Discussion Paper Number 83, Boston University, June 1982.

______, "Rivalry and the Excessive Allocation of Resources to Research," *Bell Journal of Economics*, Spring 1983, *14*, 152–65.

von Weizsacker, C. C., *Barriers to Entry*, Berlin: Springer-Verlag, 1980.

Williamson, Oliver E., "Economies as an Antitrust Defense," *American Economic Review*, March 1968, *58*, 18–36.

______, "Economies as an Antitrust Defense Revisited," *University of Pennsylvania Law Review*, January 1977, *125*, 699–723.

Wilson, Robert, "Informational Economies of Scale," *Bell Journal of Economics*, Spring 1975, *6*, 184–95.

Rationing by Waiting Lists

By COTTON M. LINDSAY AND BERNARD FEIGENBAUM*

Queues are observed to emerge in two distinct settings in economic theory. On the one hand they result as a byproduct of stochastic variation in one or both sides of markets where instantaneous adjustment of output and/or price is prohibitively costly (Arthur De Vany, 1976). Arrival times of demanders may fluctuate as with telephone services, and supply may vary stochastically as with cruising taxicabs, raising issues of optimal capacity and equilibrium price in such markets. On the other hand, queues emerge in a nonstochastic setting when price is below or above the market-clearing level due to transaction costs or external price constraints. If prices are below the market-clearing level, queues of demanders will form to ration the available supply, while in the opposite case suppliers queue for access to available customers.

This paper examines a case of the latter type in which price is too low. Conventional analysis of these queues relies on demander time devoted to standing in line to clear these markets. Even where demand is perfectly predictable and uniform, lines form and grow until the expected wait in these queues exacts a cost in time equal to the value of the goods received for the marginal demanders. Welfare and distributional implications of such a market have been developed by Yoram Barzel (1974). Many important nonstochastic queues do not rely on waiting time to clear their markets, however. Demanders of "low-cost" housing do not brave the elements for months at a time waiting for a unit to become available. Demanders of zero-price book services from public libraries do not camp in the stacks awaiting their turns at best sellers. And seekers of season tickets for the San Francisco Opera are not permanently queued before the ticket office in the lobby. The central feature of all these examples, and a feature which plays no role in the conventional theory, is the *waiting* list. In these examples and in many other cases, queuing does not necessarily imply waiting in person. Occupants of waiting lists wait, but they do so in absentia. Such waiting does not therefore impose a cost in wasted time. One is free to do whatever he wants with his time (except, of course, enjoy the services of the good sought).

If membership in waiting lists imposes no cost, we are confronted with the question of how such markets clear, and whether, in fact, they do. This paper develops the theory of waiting list queues which clear markets. The theory yields several empirical implications that differentiate it from the standard queuing by waiting time model, as well as from numerous *ad hoc* explanations for the presence and behavior of queues offered up by observers (too) close to the scene.

The central assumption of the theory is that delay in receipt of a good can lower its value to demanders. This morning's newspaper is worth more to readers this morning than this evening, and certainly its value will have fallen to nearly nothing in a week. The value of miniskirts and Nehru jackets delivered today is clearly less than the value of similar garments delivered in 1968. If one could obtain such items by costlessly adding his or her name to a list, the list itself would not diminish demand and would serve no rationing function. On the other hand, if position in the queue was linked to the date of delivery, then value and therefore the number enrolling per period will be influenced by the length of the list. It is this diminishing value rather than increasing cost

*Department of Economics, Clemson University, Clemson, SC 29631, and California State University, Northridge, CA 91330, respectively. We are indebted to Mike Maloney for the many hours he devoted to discussions involving the developments in this paper. We also wish to acknowledge helpful comments by Yoram Barzel, Chris Curran, Jerry Dwyer, Edward Greenberg, Dave Haddock, Charles Webster and a referee. A portion of this work was supported by grants from Hoffman-LaRoche, Inc. and the Emory Law and Economics Center.

of obtaining such goods that produces the convergence of quantity demanded and the quantity supplied.

A second assumption is that individual demand is unpredictable from period to period. For demands which are readily predictable (like newspapers, but unlike high fashion clothing), delay from the date of order to date of receipt need not reduce the number of demanders. As each demander of this type good may forecast his desired quantity in each future period, he may simply "order in advance" to obtain it. Lists will grow indefinitely in such cases and fail to perform a rationing function. We may infer, therefore, that queuing by lists will be employed more frequently in markets for which individual demands are episodic and unpredictable than in markets where demands may be accurately anticipated. Such a queue would be less effective in clearing the market for food, for example, since next month's nutritional needs are well known to everyone. Ordering food in advance does not lessen its value when received, hence this delay does not lessen orders for food to the point where the market cleared. We note with interest therefore that in communist countries where many goods are priced below their market-clearing levels, food is dispensed through queues of standing people, while theater tickets and holidays at Black Sea resorts are assigned from waiting lists.[1]

In Section I, we develop the theory of queuing by list, beginning with a description of the individual decision to join. This individual behavior is then aggregated by market to describe the sensitivity of the rate of joining to expected delay in delivery, the rate at which demand diminishes with delay in delivery, and other variables. A theory of supply is then sketched, and conditions for market equilibrium are developed. In Section II, the theory is tested on data from one of the largest queues in the Western world, that is, the waiting list for admission to British National Health Service (NHS) hospitals.

Our theory stands in contrast to "official" explanations of waiting lists in the National Health Service and elsewhere. The NHS maintains that waiting lists are merely backlogs. These explanations imply that the rate at which services are demanded in each period equals the rate at which they are supplied, but because of a backlog of cases, the market does not clear. Total demand in a period does not equal the period's total supply because of the holdover from previous periods, but such a deficiency does *not* indicate long-run inadequacy of resources to deal with demand. Instead it represents a backlog of cases which would and should be eliminated through concerted short-term efforts including, for example, temporarily making additional operating theaters available, diverting beds from other specialties, reducing length of stay in hospitals, performing surgery in outpatient departments, using hospitals maintained by the armed forces, etc.[2]

This official explanation has been routinely confounded in experience with the National Health Service by the repeated failure of such "short-term" remedies to produce the predicted results. On the contrary, expansion of facilities typically does not eliminate waiting lists, or even substantially reduce them.[3] The addition of new beds or staff has in many cases resulted in an increase in the numbers of patients on lists, leading some observers to reach the cynical conclusion that doctors themselves are responsible for waiting lists. According to this interpretation, additional doctors in hospitals "generate" a demand for their services by ordering more surgery and more inpatient care. Our model presents an alternative, economically more plausible, interpretation of these phenomena.

Before turning to the formal development of the theory, however, we offer some possible explanations for the existence, or more

[1] Though unpredictability in *individual* demand is a necessary condition for equilibrium in a market cleared by waiting lists, this does not imply that demand to suppliers itself is stochastic. On the contrary, we assume that market demand is regulated by the law of large numbers operating on individual demander units so that actual demand faced by suppliers is uniform from period to period.

[2] This view is forcefully stated in the Ministry of Health's *Reduction of Waiting Lists*..., (1963).

[3] See, for example, Martin Feldstein (1967, pp. 200 ff.) and A. J. Culyer and J. Cullis (1976).

correctly, the "survivability" of this form of rationing. As with the more conventional form of queuing by standing in line, queuing by list involves waste. Delay in receipt is costly, and losses due to this delay do not seem to provide gains to anyone. Yet waiting lists abound. Lists are maintained for union membership, opera tickets, public housing, club membership, and the court system, to name only a few examples of this institution here in the United States. One explanation for adoption and survival of waiting list rationing may be found in the predictability of demand just discussed.

Casual observation suggests that queuing by list is observed most often in the dispensing of services by nonproprietary organizations whose membership receives rents in the form of underpriced services, higher than equilibrium pay or the like. Membership confers these benefits, however, only during the active participation by the member himself. These rights are not alienable. They are terminated by retirement, separation, relocation, or death. However, they may be conferred on friends or family members through advantages in the process of queuing for membership. As noted above, the more predictable is one's demand, the less costly will be delay from the date of ordering to the date of delivery. One may anticipate demand and order in advance. If predictability of demand varies in the community, then those with less uncertainty will crowd out those with highly uncertain demands, securing for themselves the rationed item.

It seems plausible that unions employ this device to restrict membership to friends and families of existing members, to the disadvantage of more transient workers. Country clubs may employ the same strategy to disadvantage new arrivals in the community for the betterment of the old guard. Politicians may employ this form of rationing to reward faithful constituents with access to underpriced public housing, to the disadvantage of the more mobile poor who are less likely to vote.

The point in all these examples is not that the advantaged groups cheat by "jumping the queue," though cheating may certainly be practiced at lower cost with waiting lists than with some other forms of nonprice rationing. Rather, we maintain that these lists serve as components of *rationing* processes. Those disadvantaged by the process do not merely suffer greater delay in delivery, as they might, were queue jumping practiced on a large scale. They get none at all. The mechanics of this rationing process are presented in Section I below. Variation in the predictability of demand is one characteristic confering advantages in the rationing process, but there are others. In order to simplify the presentation of that theory as it applies to these other characteristics, we therefore assume sufficient uncertainty concerning future demand to make signing up in anticipation of future demand uneconomical for all.

I

A. *Individual Joining*

A person will join a waiting list if the present value of the good when delivered exceeds the cost of joining the queue. To simplify, we assume that upon arriving at the top of the waiting list, each demander is entitled to purchase a fixed amount of the good at a price (possibly zero) below the market-clearing price. The value of the rights obtained by joining such a list will therefore depend on price. On the basis of the assumption stated earlier, the value of these rights will also depend on the delay in receipt of the good.

Delay affects the present value of these rights in two ways. If the interest rate is positive, the right to consume the good in the future is discounted relative to its present exchange value due simply to market-expressed time preference. Delay may also affect the present value of consumption for reasons that have nothing to do with interperiod exchange rates. The timing of delivery may affect a good's value due to fashion, circumstance, location, health, or whim. If one wishes to return home today, a plane reservation today is worth a great deal more than a reservation tomorrow, and exceeds the value of a reservation next week by an even larger margin. The effect of both the discount rate and diminishing demand may be expressed in exponential form. Rather than carry them both through the analysis, we will express their combined effect by an exponential demand decay rate g. As the delays with which we deal empirically rarely exceed several months in duration, and the chief thrust of the analysis is to predict the influence of *differences* in decay rates in different queues, the discount rate component of g plays little role and may be ignored.

The present value of the rights obtained from joining a waiting list may in this case be expressed as the product of their current value v (dependent on a vector of unknown attributes $\tilde{u}$ and the delivery price p), and an exponential in the decay rate g and the expected delay in delivery t. For the ith person, this value is given by

$$v_i(\tilde{u}, p, g, t) = v_i(\tilde{u}, p)\cdot e^{-gt}.$$

The cost of joining the list includes any costs incurred to qualify for joining other than the purchase price. These include costs of taking examinations, obtaining approvals and referrals, and such transactions costs as expenditures for transportation, legal advice, and market information. Let the value of all these costs incurred by the ith individual be c_i.

For given values of c_i, v_i, and g, equality of the value and cost of joining a list obtains for a single value of t. Let this solution value of t for the ith individual be $\hat{t}_i$. Figure 1 depicts this relationship between benefits, costs, and time for such a queue. If the value at the date of ordering is v_1, the decay rate is g_1 and joining costs are c, the critical delay for joining the list is given by $\hat{t}_1$. If the expected wait in the queue is less than $\hat{t}_1$, the benefits of joining exceed the cost and the individual will join. If the expected wait is greater than $\hat{t}_1$, the benefits will be less than the cost, and the individual will "balk," that is, decline to join.

As v and c vary among demanders, this critical value of delay $\hat{t}$ itself will vary. The decision of the ith person to join a queue is determined by the relation of $\hat{t}$ to the expected delay t. For those who join, $\hat{t}$ must be equal to or greater than expected delay t. For those who do not $\hat{t} < t$. We shall concern ourselves with conditions for the *marginal joiner*, that is, the demander for whom $\hat{t} = t$. As expected delay varies, equality of $\hat{t}$ and t implies that $v\cdot e^{-gt} = c$ must hold at the margin. *Ceteris paribus* changes therefore imply the following relationships for marginal joiners:

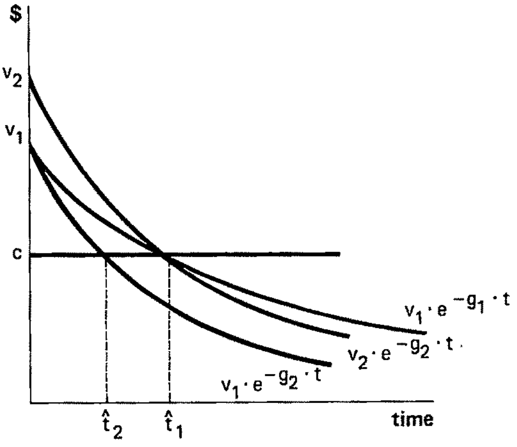


FIGURE 1

$$\partial v/\partial g = v\cdot t > 0 \quad \text{and} \quad \partial v/\partial t = v\cdot g > 0.$$

These relationships may be illustrated with reference to Figure 1. An increase in decay rate from g_1 to g_2 will shorten $\hat{t}$ for the the individual who places a value of v_1 on the good received from the queuing process. He will no longer join the queue. At the original wait of $t = \hat{t}_1$, the marginal joiner will now be the chooser who values the good at v_2, which is higher than v_1. Similarly, holding the decay rate constant at g_2 while increasing the expected wait from $t = \hat{t}_2$ to $t = \hat{t}_1$ also increases the value placed on the good by the marginal joiner from v_1 to v_2.

B. *The Rate of Joining*

For our application as well as most occasions in which waiting lists are employed, we may treat the purchase price as given and the cost of joining as uniform across persons. Variation in $\hat{t}$ is thus attributable to the decay rate and to variation in the vector of consumer attributes $\tilde{u}$ which varies v. We wish to describe the relationship between expected wait for delivery and the rate at which people will join such a queue. This clearly involves defining the relationship between the distributions of people with varying v and g and the distribution of $\hat{t}$ in the population. We will do this by first assuming that everyone in a given queue has the same decay rate, then observe the effect of changes in g on particular properties of the queue.

If decay rate is unchanging, the only factor giving rise to variation in $\hat{t}$ among the population is v. Let us assume that the distribution of v is described by the frequency distribution $f(v)$. We assume that $f(v)$ is continuous and that the range of v is finite, that is, $0 \leq v \leq \bar{v}$. An expected wait of t_1 will imply that joining the list will be conditional on valuing the good by some amount v_1 or more. We may thus describe the number who join the queue per period as a function of v and population N:

$$(1) \quad h(v) = N \int_{v}^{\bar{v}} f(v)\, dv = N[1 - F(v)],$$

which may be related to $\hat{t}$ by merely substituting for v,

$$j(\hat{t}) = N[1 - F(c \cdot e^{g\hat{t}})].$$

The function $j(\hat{t})$ is the number of people for whom the critical delay has a value of $\hat{t}$ or more, so the function $j(t)$ is the number who would join the queue at an expected wait of $t = \hat{t}$, that is,

$$j(t) = N[1 - F(c \cdot e^{gt})].$$

This joining function forms one structural component of a system which simultaneously determines the expected delay, the number in the queue, the rate of joining, and the rate of service. The $j(t)$ intercept of this function is determined by evaluating it at $t = 0$:

$$(2) \qquad j(0) = N - N \cdot F(c).$$

The interpretation of this intercept is straightforward. It is the number of demanders of this good who value it by more than the cost of joining the queue. We thus refer to this intercept as the number of *potential joiners*, since this is the number of people willing to sign up at no expected delay in delivery. The slope of this joining function is given by

$$\begin{aligned} dj/dt &= -N \cdot f(v) \cdot \partial v/\partial t \\ &= -N \cdot f(v) \cdot g \cdot v \end{aligned}$$

which must be negative.

The shape of such a joining function depends obviously on the distribution of v among demanders just as the shapes of conventional demand curves depend on the distributions of demanders willing to pay given money prices under such circumstances.[4] One of our chief concerns, however, is the influence of the decay rate on the position of the joining function. It can be shown that an increase in the decay rate will rotate the joining function predictably around its joiner intercept. Note first that changes in the decay rate will *not* affect this intercept of the joining function. At $t = 0$ the intercept term, as given by expression (2), is unaffected by variation in g. For any positive delay, however, an increase in the decay rate reduces the rate of joining. That is,

$$\partial j/\partial g = -N \cdot f(v) \cdot t \cdot v < 0.$$

There remains one important extension before this theory of waiting list joining may be usefully employed. If we are to compare queues in different communities or for different goods, we must somehow normalize for the number of potential joiners in each queue. Some communities are larger than others, and some goods more popular than others within given communities. The number of potential joiners may therefore vary from list to list. In terms of hospital waiting lists, for example, we expect more joiners in more

[4] The second derivative of the joining function is

$$d^2_j/dt^2 = -N \cdot g^2 \cdot v[f'(v) \cdot v + f(v)]$$

which cannot be signed without some assumption about $f'(v)$. For the case of a rectangular distribution where $f'(v) = 0$, d^2j/dt^2 is unambiguously negative, and the function is concave. For "bell-shaped" distributions, concave and convex portions of the curve are possible.

populous regions than in sparsely populated areas and, given the differences in the incidence of diseases, more joiners in the queues for treatment of tuberculosis than in queues for encephalitis. Furthermore, we typically have only the sketchiest information on the number of potential joiners, particularly when we wish to compare queues within a community. The fact that more people are observed to join one queue than another may simply imply, as we have seen, that the expected delay in the delivery may differ between queues. As expression (2) above provides us with an estimate of the number of potential joiners, however, we may use this information to normalize the parameter estimates of this function.

We assume that this variation in $f(v)$ from market to market is fully described by a scalar K_j such that the frequency of the qth queue at any v is merely K_q times the frequency in some numeraire queue (where $K=1$), that is,

$$f_q(v)=K_q\cdot f(v).$$

The joining function of the qth queue may then be represented

$$j_q(t)=K_q\cdot N_q[1-F(c\cdot e^{g\cdot t})]$$

with intercept

$$j_q(0)=K_q\cdot N_q[1-F(c)]$$

and slope

$$dj_q/dt=-K_q\cdot N_q\cdot f(v)\cdot g\cdot v.$$

Comparison of estimates of the structure of alternative joining functions is ambiguous, as both intercept and slope will contain different N and K. We do not know, in other words, whether a high coefficient on expected wait in any particular case reflects higher decay rates or simply more potential joiners. Individual estimates may nevertheless be normalized by noting that the ratio of constant terms of two such estimates for the qth and rth queue, respectively, is itself the ratio of the potential joiners in each queue, providing a ratio of the unknown parameters,

$$\frac{K_q\cdot N_q[1-F(c)]}{K_r\cdot N_r[1-F(c)]}=\frac{K_q\cdot N_q}{K_r\cdot N_r}.$$

The slopes of two such joining functions may thus be made comparable by dividing the estimated slope of one by the ratio of the estimated intercepts, for example, dividing the slope of the rth equation by the ratio $K_q\cdot N_q/K_r\cdot N_r$. Alternatively, each observation of the rate of joining the rth queue may be deflated by this ratio, and a normalized joining function may be estimated directly.

A description of the response of people joining waiting lists to variation in expected delay in delivery is not a complete theory of waiting lists. The expected delay itself must be determined, and this depends not only on the rate of joining, but on the rate of supply. The equilibrium wait is that which clears the market of the available supply.

C. *The Rate of Supply*

It is not our purpose here to develop the supply side of this model beyond recognizing that, whatever other influences affect the quantity of these goods supplied, waiting lists *may* exercise an additional influence on supply. We hypothesize, in other words, that supply at time h depends upon a vector of unknown determinants $\tilde{w}$ and is positively affected by the waiting time for that good: $s_h(\tilde{w},t)$, subject to $\partial s_h/\partial t>0$.

D. *The Market Determination of t*

The membership of each queue will consist of those who have previously joined who have not yet reached the top of the list. At time h that will represent the sum of demands previously unsatisfied, net of exits from the queue (which we ignore), that is,

$$Q_h=\sum_{k=0}^{\infty}(j_{h-k}-s_{h-k}).$$

The rate of change in the numbers in the queue in any period h is therefore given by $\dot{Q}_h=j_h(t_h)-s_h(t_h)$.

The expected wait in period h is the total number in the queue divided by the service rate, that is, $t_h = Q_h / s_h$, and reaches an equilibrium t^* when $t_h = t_{h+1}$. This will occur clearly when $\dot{Q} = 0$ implying a solution at t^* and Q^* such that $j(t^*) = s(t^*)$, and $Q^* = j(t^*) \cdot t^*$. As $dj_h/dt < 0$ and $ds_h/dt > 0$, expected wait will converge to this equilibrium. For values of $t < t^*$ is must be the case that $\dot{Q} > 0$, and for values of $t > t^*$, $\dot{Q} < 0$.

The wait in such a list functions, therefore, is very much like a price. Instead of clearing the market by raising the cost of obtaining the good, however, waiting time clears the market by making it less valuable. If a good is distributed to a population with varying v and g, those demanders with high values and low decay factors will crowd out demanders with lower v and higher g.

The description of this equilibrium permits us to perform some comparative statics involving queue length, expected wait, and the underlying variables in the model. These are illustrated in Figure 2. Let the initial joining function be as indicated by $j_1(t)$ and the initial supply function by $s_1(t)$ yielding equilibrium waiting time t_1^*. Clearly an autonomous permanent increase in supply to $s_2(t)$ decreases equilibrium waiting time to t_3^*.

The effects of changes in the decay rate g play an important role in our empirical analysis. Clearly an across-the-boards increase in this decay rate will shift the joining function $j(t)$ in the manner described in the previous section to a position such as $j_2(t)$ diminishing the equilibrium wait from, say, t_1^* to t_2^*. In our empirical analysis, however, we are not concerned with universal changes in the decay rate, but rather in the analysis of different queues in which decay rates are homogeneous among members of each queue but which differ between queues. Indeed, it is our contention that much of the variation in waiting lists for hospitalization in the British NHS can be explained in terms of the differing decay rates for different hospitalizable conditions.

Our ability to make such comparative static predictions depends critically on the theoretically implied structure of the joining function. Our empirical efforts will therefore focus on the estimation of this function. Before turning to this estimation, however, we will return to the theory of these joining functions to derive an expression for the elasticity of joining with respect to expected wait in the queue. Such an expression has two valuable uses in this paper: first, it permits us to hypothesize a priori relationships between elasticities of joining among different queues that are then used in the empirical section. Second, it permits us to explain an anomaly often observed in connection with NHS queues. Increases in the service rate are occasionally accompanied by increases in the number of patients joining these queues. Such behavior is clearly inconsistent with the official "backlog" explanations of these waiting lists. It is easy to show, however, that this is precisely the result expected from our model when this elasticity is greater than one in absolute value.

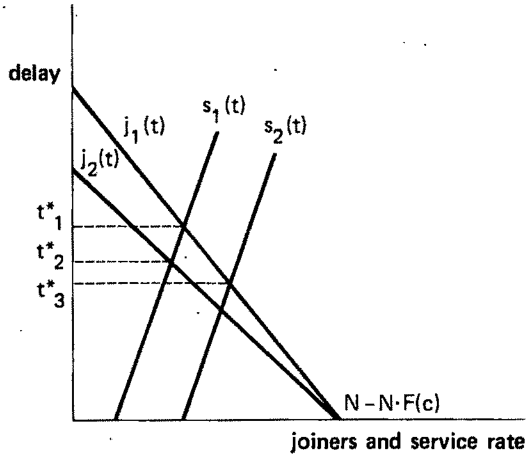


FIGURE 2

E. *The Elasticity of Joining with Respect to Delay*

Let us return to our more restrictive assumptions concerning the composition of a queue. Where the decay rate is uniform, and the supply price of the delivered good is the same for everyone, value alone gives rise to variation in $\hat{t}$ and thus to the rate of joining. Expression (1) above identifies this relationship $h(v)$. The elasticity of joining with respect to wait is $\varepsilon_t = t/j \cdot dj/dt$. The term dj/dt may be decomposed into $dh/dv \cdot dv/dt$ where dv/dt is the change in the value placed on the good by "marginal joiners" per unit change in expected delay. This elasticity may thus be rewritten

$$(3) \qquad \varepsilon_t = \frac{t}{j} \cdot \frac{dh}{dv} \cdot \frac{dv}{dt} = \frac{t}{j} \cdot \frac{dh}{dv} \cdot g \cdot v.$$

The elasticity of joining with respect to value placed on the good by the marginal joiners is $\eta_v = v/h \cdot dh/dv$.

Close examination of expression (1) reveals that η_v will under certain circumstances be the demand price elasticity of the good distributed. This will be the case where the delivery price of the good p is equal to zero, and each demander is restricted to purchasing a fixed quantity of the good (no more, no less) per appearance at the head of the queue.[5]

The expression for demand elasticity may be solved for the slope term $dh/dv = \eta_v \cdot h/v$ which substituted into expression (3) gives the elasticity of joining the waiting list with respect to delay as dependent on expected wait itself, the elasticity of demand (in terms of clearing prices) and the decay rate

$$\varepsilon_t = t \cdot \eta_v \cdot g.$$

For any given expected wait, in other words, the responsiveness of those joining waiting lists to changes in this wait will vary positively with the demand elasticity and the decay rate. Note that, since t and g are both positive, the elasticity of joining with respect to delay will have the same sign (negative) as the demand price elasticity of the good.

F. *The Number in the Queue*

It was pointed out above that in equilibrium, the number in the queue Q would be equal to the joining rate $j(t)$ times the expected wait t. We have also shown that an increase in supply will unambiguously reduce the equilibrium wait. A change in the expected wait should therefore have the following effect on the number in the queue:

$$dQ/dt = j(t) \cdot (1 + \varepsilon_t),$$

hence supply shifts will have the following effects

$$\frac{dQ}{ds} = \frac{dQ}{dt} \cdot \frac{dt}{ds} = \frac{dt}{ds} \cdot j(t) \cdot (1 + \varepsilon_t)$$

which may be grouped according to elasticity

$$dQ/ds \gtreqless 0 \quad \text{as} \quad |\varepsilon_t| \gtreqless 1.$$

In the introduction we pointed out that the observation of "perverse" behavior by waiting lists in response to changes in capacity has led some observers of English NHS queues to conclude that the hospital consultants themselves were responsible. We see now that it is not necessary to impugn the motives of doctors in the system to explain how increases in capacity may lead to longer waiting lists. On the contrary such a result is expected in certain cases. To be sure, the expected wait on these longer lists will be of shorter duration. But so long as the elasticity of joining with respect to expected wait has an absolute value greater than one, an increase in supply will result in a waiting list containing *more* names![6]

[5] Under these circumstances the demand function in terms of market-clearing prices p will itself be a cumulative frequency as this clearing price falls from the maximum price anyone will pay $\bar{p}$ to zero, i.e.,

$$h(p) = N\int_p^{\bar{p}} f(p)\, dp = N[F(\bar{p}) - F(p)],$$

where the delivery price is itself zero, $v = p$, and $h(v) = h(p)$. Thus,

$$\eta_v = v/h \cdot dh/dv = p/h \cdot dh/dp.$$

[6] Judith Mann (1974, pp. 60 ff.) has produced a related result in a completely different model. Queues form in her model due to stochastic characteristics of demand. *Service* time is endogenous, and demand varies inversely with the duration of service time but is insensitive to delay in these queues. A decrease in service time increases the rate at which orders are processed, a result that is interpreted in standard operations research models as lowering the average queue length. In Mann's model, however, this reduction in service time increases the "quality" of the product, leading to an increase in orders. Moreover, the nature of the equilibrium derived

II

The empirical focus of this paper is the queue for hospitalization under the National Health Service in Great Britain. Waiting lists have been associated with NHS since its inception in 1948. The process for joining such a queue for nonemergency cases follows an invariant routine. Patients present their symptoms to a general practitioner. If the GP feels that the services of a consultant (specialist) are required, he refers the patient to the appropriate unit in the district hospital. Patients may visit these consultants only on a referral basis. If the consultant feels that hospitalization is warranted, he adds the patient to the list of those queued for that particular service, and the patient begins his wait.

Once entered on the list, the patient returns to his private pursuits and bears no further costs of being in the queue. The costs of such queuing to the patients are therefore exclusively "up front." An in-kind, standing-in-line time cost cannot be counted on to clear such a market, because these costs do not grow as the queue grows longer. Yet the market does seem to clear. In the absence of some equilibrating mechanism, we should expect these queues to grow without limit by the amount of the excess demand in each period, but that is not observed. The NHS waiting lists have grown over time, but certainly not at this rate. After one year of operation, waiting lists for hospitals contained 460,000 names. Twenty-six years later in 1976 this had grown to only 607,000.

Clearly the missing element here is the effect of time on demand. In the limiting case, some of those who joined the queue will have recovered or moved away, or even died awaiting treatment. It is our maintained hypothesis, however, that the effect of time on this process is not limited to removing from the list those who have left the queue for one of these reasons. On the contrary, we hypothesize that the expected wait itself reduces the attractiveness of joining in the first place, and that this influence operates in a way that is consistent with the theory of waiting lists that we have just presented.

That theory has suggested several empirical propositions which serve both to test the theory and to add to our understanding of the allocative results of NHS queues. Most are concerned with the structure of the joining function for these queues. Our theory implies that the rate of joining should be inversely related to 1) the expected delay, 2) the decay rate, and 3) the cost of joining.

The rate of joining will be positively related to 4) the value of the services provided. The model suggests, furthermore, that 5) membership in the queue will be positively related to the rate of service where $|\varepsilon_t|>1$ and negatively related to membership in the queue where $|\varepsilon_t|<1$. We lack data on the cost of joining, so relationship 3 cannot be empirically explored here. Our statistical work must therefore focus on the remaining variables, that is, the responsiveness of joining to the expected delay, value, the decay rate, and the elasticity of joining with respect to the expected delay.

A. *The Data Base*

We are interested in relating the rate at which demanders of hospital services join waiting lists to expected delay and the decay rate of demand for these services. The NHS is divided into fourteen administrative regions, and data on both the mean waiting time and the number of discharges is reported annually for each ICDA disease category by region.[7] We have therefore chosen to treat the waiting list for each condition in

implies that "demand is sufficiently elastic with respect to time changes that the arrival rate will increase by a greater percentage than service time is decreased, leading to an increase in both the mean order stock and the mean queue length" (p. 61). In our model service time is constant, hence there is no scope for the effect described by Mann. Output changes are achieved through an increase in the rate of production of constant duration services. As delay in receipt performs a rationing function, an increase in the service rate cannot result in an increased delay, though it may increase the number of demanders in the queue as shown.

[7] Department of Health and Social Security (DHSS) 1978. Data on population, hospital beds, and doctors by region were taken from DHSS (1975). Data will be supplied upon request.

each hospital region as a separate queue for estimation purposes. Data for the year 1974 were employed.

This treatment was dictated by the form of the data, and has two short-comings which we acknowledge but are powerless to correct. In the first place, waiting lists are maintained by hospital rather than region. The use of data aggregated to this level masks the intraregional variability in waiting times and the other variables in our analysis. On the other hand, demanders are not restricted to a single hospital, but may "shop" among alternatives in their region for hospitals offering the shortest wait. Intraregion variability in the wait for a particular type of hospitalization is therefore assumed to be small relative to interregional variation. Second, separate queues for admission are not explicitly maintained in each hospital for every procedure and condition. The division of beds controlled by each specialist depends on the size of the hospital and its staff, and even in the largest, a single specialist assigns patients to beds for a variety of conditions. Still, the extreme variation in these waiting times from condition to condition is inconsistent with the view that all patients are admitted from a small number of waiting lists. On the contrary, this wide variation in waiting times suggests that separate queues are implicitly maintained, perhaps by the practice of admitting cases of different types in roughly fixed proportions.

An additional difficulty presents itself in connection with our variable, the decay rate. Lacking any objective measure of this decay rate for individual hospitalizable conditions, we choose to assign decay rates by grouping them. With the assistance of a practicing physician,[8] we sorted all diseases and other hospitalizable conditions into three groups:

1) Category I diseases are those with what appeared to be high demand decay rates. These were nonemergency cases, typically susceptible to drug therapy for which alternatives to hospital care were available, but cases that respond to treatment and are controlled within a reasonable time in most instances, even if hospitalization is not provided.

2) Category II conditions, on the other hand, were nonemergency cases such as hernia or cataracts that do not grow worse with delays in treatment, but for which no alternative to hospital based therapy is available. These were treated as low decay-rate conditions.

3) Category III conditions were those which rapidly grow more serious over time. These conditions might be said to have negative decay rates in the sense that demand increases rather than subsides with the passage of time.

Table 1 presents those classes of cases for which data were used, grouped by decay category and identified by ICDA code number. Certain conditions and diseases were deleted from our categorization due to missing observations or ambiguity in assignment to a category.

B. *The Empirical Specification*

The theory suggests the following relationship between the rate of joining, expected wait, decay rate, and the value of service:

$$j = j(t, g, v),$$

$$\partial j/\partial t < 0, \quad \partial j/\partial g < 0, \quad \partial j/\partial v > 0.$$

Indeed, the theory implies that the effect of changes in the decay rate is to rotate the function with respect to time around its joiner intercept. If linear estimation of this function is adequate, the following specification is implied:

$$(4) \quad j = a_0 + a_1 \cdot t + a_2 \cdot g \cdot t + a_3 \cdot v + u$$

in which coefficients are predicted to have the signs indicated: $a_1 < 0$, $a_2 < 0$, and $a_3 > 0$.

Observable variation in decay rate is limited to the grouping described in the preceding section. A dummy variable with a value of one therefore indicates inclusion in the high decay rate, Category I, while a zero value was inserted for the low decay rate,

[8] We wish to acknowledge the assistance of Leopoldo Hernandez, M.D., in the assignment of diseases to these three categories.

TABLE 1—DISEASE CATEGORIES GROUPED ON THE BASIS OF "DECAY RATE"

Codes	
Category I: Diseases with Fast Decay Rates	
A1–44	Infective and parasitic diseases
A62–63	Goitre and thyrotoxicosis
A74	Epilepsy
A75	Inflammatory disease of eye
A97	Diseases of teeth and supporting structures
A98	Peptic ulcer (excluding gastro-jejunal ulcer)
A136	Senility (without mention of psychosis)
AN141–42	Dislocation (without fracture) and sprains of joints and muscles
Category II: Slowly Decaying Diseases	
A61	Benign neoplasms and enoplasms of unspecified nature
A79b	Strabismus
A76	Cataract
A78,79d	Diseases of the ear and mastoid process
A88a	Varicose veins of lower extremities
A88b	Hemorrhoids
A94	Hypertrophy of tonsils and adenoids
A101a	Hernia
A109,111b–c	Male genital disorders
A110,111d–f	Diseases of breast and female genital system
A121	Arthritis and spondylitis
A125a	Internal derangement of joint
A125b	Displacement of intervertebral disc
A112–24,125c	Other musculoskeletal and connective tissue disorders
Category III:	
A45–58	All malignant neoplasms
A59–60	Neoplasms of lymphatic and haematopoietic tissues
A64	Diabetes mellitus
A73	Multiple sclerosis
A80–81	Rheumatic fever and rheumatic heart disease
A82	Hypertensive disease
A895	Cerebrovascular disease
A87	Venous thrombosis and embolism
A91–92	Pneumonia
A100	Appendicitis
A103	Cholelithiasis and cholecystitis
A107	Infections of kidney
A108	Calculus of urinary system
AN138	Fracture of skull and face
AN140a	Fracture of neck or femur
AN104b	Fracture of other and unspecified parts of femur

Category II. It is the practice of the NHS to treat Category III conditions differently from cases in the other two categories. These emergencies are understandably moved to the head of the queue and do not follow the process outlined in our theory. Cases of this kind are therefore inframarginal to the determination of equilibrium waiting time and have no influence on results predicted for the remaining categories. Category III observations were therefore dropped from our sample. This leaves us with 22 conditions observed in 14 regions for a total of 308 observations.

Three problems remain concerning the estimation of this joining function. First, we have no data on numbers actually joining each queue in each period. Our model suggests, however, that in equilibrium the rate of joining will just equal the rate of output s, and we have data on service rates per period. Second, as we have hypothesized that the rate of supply may also be influenced by delay, ordinary least squares (OLS) estimates of these coefficients may be inconsistent. To remedy this, we employ predictors of t that are uncorrelated with the disturbance term u. The predicted values of t are used in the joining equation to obtain unbiased estimates of its structure.[9] Delay is scaled in weeks.

The third problem concerns variation in the number of potential joiners. Recall that

[9] Predictors of t were obtained by regressing mean waiting time (MWT) for each condition by region on hospital beds available by region per 1,000 population (BAP), number of MDs on hospital staffs per 1,000 population (MD), and joiner equation variables including a dummy for disease category (D) and ($NEDO$) measuring the proportion of the regional population 65 years and older. This yielded the following estimate (the t-ratio are shown in parentheses):

$$\begin{array}{ll} MWT = \underset{(6.84)}{28.7} - \underset{(-3.46)}{1.28 \cdot BAP} - \underset{(-2.21)}{9.60 \cdot MD} & \\ \qquad + \underset{(1.35)}{.33 \cdot NEDO} - \underset{(-10.15)}{7.00 \cdot D}; & R^2 = .30. \end{array}$$

When other age distribution variables were used instead of $NEDO$ in the joining equation estimates, they were also substituted here. As expected, this had a negligible effect on predicted delay.

TABLE 2—INSTRUMENTAL ESTIMATES OF THE JOINING FUNCTION, UNADJUSTED AND ADJUSTED FOR POTENTIAL JOINER GROUPS FOR EACH CATEGORY[a]

		Adjusted		Adjusted		Adjusted
Variable	(1)	(2)	(3)	(4)	(5)	(6)
C_1	3.920 (2.34)	-	4.414 (1.13)	-	4.542 (3.26)	-
C_2	1.803 (1.24)	-	2.163 (0.54)	—	2.173 (2.04)	-
C	-	4.252 (2.72)	-	5.037 (1.12)	-	4.542 (4.55)
t_1	−.0915 (−1.18)	-	−.1002 (−1.09)	-	−.1062 (−1.26)	-
t_2	−.0991 (−0.92)	-	−.1000 (−0.86)	-	−.0978 (−0.88)	-
t	-	−.1076 (−1.78)	-	−.0969 (−1.43)	-	−.1062 (−1.74)
$g \cdot t$	-	−.089 (−1.81)	-	−.0983 (−1.82)	-	−.0983 (−1.98)
V_1	.0273 (0.36)	.0225 (0.24)	-	-	-	-
V_2	-	-	.0008 (0.01)	−.0173 (−0.13)	-	-

Note: t_1 = mean delay, Category II; t_2 = mean delay, Category I; t = mean delay, all observations; $g \cdot t$ = delay times decay dummy; V_1 = percentage of population over age 64; V_2 = percentage of population over 64 or under 15.

[a] t-ratios are shown in parentheses.

this difference will affect both the estimated constant term and the slope of the joining function. We have chosen to control for this difference using the technique suggested in the theoretical section. It was determined there that the constant term in an estimate of the joining function for a single queue yields the number of potential joiners. By separately estimating regressions for each queue and identifying these constants, we may "deflate" the dependent variable of one queue by the ratio of the constant terms making the two queues comparable.

Columns 1, 3, and 5 of Table 2 present unadjusted estimates of the coefficients on delay. The dependent variable in these estimates is cases treated per 1,000 population which proxies the rate of joining as noted above. Two additional demand variables were included to test the hypothesized sign of a_3. These are V_1, the proportion of the regional population over age 64, and V_2, the proportion of the population either over 64 or under 15 years. Both were included based on the assumption that these age groups are disproportional users of hospital care and should value it more highly on average. Neither of these variables were found to have a statistically significant influence on the rate of joining. Columns 5 and 6 report estimates of this joining function with no demand shifter variable.

The influence of delay and the decay rate is best assessed in columns 2, 4, and 6. In these equations, observations of the dependent variable for Category I cases were deflated by the ratio of the estimated constant terms C_1 and C_2 from columns 1, 3, and 5. The regressions were then forced through a single intercept.[10]

Our theory implies a negative relation between expected delay and the rate of joining. It also implies a negative coefficient on the interaction of decay rate and the effect of delay. Both implications find some support in the estimates reported in columns 2, 4, and 6 of Table 2. All coefficients on the

[10] A test on this restriction revealed no significant effect on the estimate.

delay instruments (t) in these three equations have the predicted sign, and all are significantly different from zero at the 10 percent level or lower. We may interpret the coefficients on $g \cdot t$ to suggest that slopes on the adjusted joining functions for high decay-rate diseases are more than 80 percent larger than those on low decay-rate conditions. The hypothesis of no difference in decay rates is rejected at the 5 percent level or better in all specifications.

TABLE 3—SUPPLY EQUATION[a]

Variable	(1)	(2)	(3)
C	−5.796	−5.860	−5.880
	(−3.98)	(−4.01)	(−4.02)
MD	4.279	4.302	4.310
	(2.87)	(2.89)	(2.89)
BAP	.2925	.2948	.2956
	(2.37)	(2.39)	(2.40)
t	.2182	.2203	.2210
	(6.52)	(6.54)	(6.55)

[a] See Table 2.

Table 3 contains estimates of the supply function in our system. The dependent variable in this case is again cases treated per 1,000 population. Not surprisingly, beds available per capita (*BAP*) and doctors per capita (*MD*) enter positively and significantly. More interestingly, as hypothesized above, output also seems responsive to expected delay. As Table 2 employs three separate instruments for delay, the supply equation was estimated for each. Coefficient estimates for delay are quite stable across equations. The longer the expected wait per condition, the higher is the rate of output.[11] This finding is particularly noteworthy in view of the sign change on this coefficient from Table 2 results to those of Table 3.

Elasticities have been computed from joining and supply equation coefficients. Elasticity of joining with respect to delay ranged from −.55 to −.64 for low decay-rate conditions. Higher decay-rate conditions were slightly more responsive according to this measure ranging from −.65 to −.70. The elasticity of the supply response to delay is also noteworthy. We placed this parameter in the neighborhood of 1.3. This estimate suggests the possibility of severe specification error for those who would estimate production functions for NHS hospitals without including a delay variable.

We conclude by developing estimates of the demand decay rates themselves for each of the two categories of illness used in the empirical section. Recall that the elasticity of joining for the ith queue may be expressed as a function of expected delay, the elasticity of demand and the decay rate, that is, $\varepsilon_t^i = t^i \cdot \eta_v \cdot g^i$. Solving for the decay rate, gives $g^i = \varepsilon_t^i / t^i \cdot \eta_v$, which may be evaluated for each category. Mean expected delay for Category I conditions was lower than that for Category II, as might have been predicted. These delays were 9.39 weeks and 16.39 weeks, respectively. Given the similarity of the delay elasticities, a higher decay rate for Category I conditions is assured for any reasonable value of η_v. There are no studies of hospital demand in Great Britain, but American and Canadian studies seem to agree that demand in these countries is very inelastic at low effective (net of insurance) prices.[12] As the money price of hospital care is zero under the NHS, this is the price range relevant to our current analysis. Under the assumption that this elasticity is −.3, the decay rate for Category I conditions is estimated to be about .24 while the decay rate for Category II conditions is .12.[13]

[11] These results confirm those of Feldstein, pp. 151–53 who found capacity utilization responsive to waiting list pressure using simple *OLS* estimates.

[12] Richard Rosett and Lien-fu Huang (1973) estimate the demand price elasticity for hospital care for a range of copayment rates and find that it rises from −.35 at a copayment rate of .2 to −1.5 at a copayment rate of .8. M. L. Barer, R. G. Evans and G. L. Stoddart (1979, pp. 25–46) survey results of a number of studies in Canada and the United States and conclude that the elasticity may be even lower for prices in this range.

[13] Recall that our variable g includes not only the effect of delay due to the factors on which we have focussed here, but also the discount rate, i.e., $g = g' + r$ where r is the discount rate, and g' is the rate at which the usefulness of the good diminishes over time. If the discount rate for these demanders was 10 percent (discount rates for persons requiring hospitalization will not be low), estimates of g' are .14 and .02, respectively.

III. Concluding Comments

This paper has developed a model in which waiting list queues function as a rationing process. As membership in such a queue itself imposes no cost, waiting lists may ration only through the influence of delay on the value of the service delivered. This finding suggests two important conclusions. Where markets or queuing *nodes* serve households for whom demand diminishes at different rates, rationing will occur on the basis of decay rates as well as value. It follows, therefore, that some obtain the rationed good who value it less than others discouraged from joining waiting lists. To the inefficiency occasioned by delay in delivery must therefore be added an additional social loss attributable to the fact that the distribution of goods among competing demanders by waiting list itself is inefficient.[14]

On the other hand, to the extent that individual nodes serve homogeneous demanders, our theory permits us to make comparative static predictions about the response of such queues to changing market conditions, and to test these implications with empirical data. This latter implication has formed the chief focus of the present paper. We have used the theory developed here to model the structure of waiting lists for British NHS hospital care with encouraging results. The theory implies that the rate of joining will be negatively related to expected delay in supply and to the rate at which demand diminishes over time. Supply, on the other hand, was hypothesized to respond positively to expected delay. Both of these implications are supported in the statistical estimates of the two structural equations. The theory also permits us to estimate the rates at which demand decays over time. Decay rates for the queues examined were both found to be positive, and hypotheses concerning their relative magnitudes were also confirmed.

[14] This argument is developed in Lindsay (1980).

REFERENCES

Barer, M. L., Evans, R. G. and Stoddart, G. L., "Controlling Health Care Costs by Direct Charges to Patients: Snare or Delusion?," Occasional Paper 10, Ontario Economic Council, Toronto, 1979.

Barzel, Yoram, "A Theory of Rationing by Waiting," *Journal of Law and Economics*, April 1974, *17*, 73–96.

Culyer, A. J. and Cullis, J., "Some Economics of Hospital Waiting Lists in the NHS," *Journal of Social Policy*, July 1976, 5, 239–64.

De Vany, Arthur, "Uncertainty, Waiting Time, and Capacity Utilization: A Stochastic Theory of Product Quality," *Journal of Political Economy*, June 1976, *84*, 523–42.

Feldstein, Martin S., *Economic Analysis for Health Service Efficiency*, Amsterdam: North-Holland, 1967.

Lindsay, C. M. *National Health Issues: The British Experience*. Nutley: Roche Laboratories, 1980.

Mann, Judith K., "The Equilibrium of Service Firms," unpublished doctoral dissertation, University of California-Los Angeles, 1974.

Rosett, Richard and Huang, Lien-fu, "The Effect of Health Insurance on the Demand for Medical Care," *Journal of Political Economy*, March/April 1973, *81*, 281–305.

Department of Health and Social Security, *Health and Social Service Statistics for England and Wales, 1974*, London: HMSO, 1975.

______, *1974 Hospital In-patient Enquiry*, London: HMSO, 1978.

Ministry of Health, *Reduction of Waiting Lists, Surgical and General*, Hm(63)22. London: HMSO, 1963.

The Effects of Expectations on Union Wages

By ROGER T. KAUFMAN AND GEOFFREY WOGLOM*

A major goal of macroeconomic policy is a lower rate of inflation accompanied by full employment. To achieve this goal it is necessary to reduce inflationary expectations and the rate of growth of wages. There are major questions, however, about the speed with which inflationary expectations change and the sensitivity of wages to slack labor market conditions. As a result, economists disagree about the costs of reducing inflation by restraining aggregate demand (Edward Gramlich, 1979, and William Fellner, 1979). Conclusive empirical evidence on this issue is difficult to gather because of problems in estimating macroeconomic models when expectations must be measured indirectly from past values of economic variables (for example, when using the adaptive expectations model; see Robert Lucas, 1976).

In this paper we address this issue by analyzing the effects of different measures of expectations on union wages using microeconomic data. Two of our measures, expected inflation and expected labor market conditions, are the forecasts of a prominent commercial forecasting firm. Since these are direct measures, they should avoid the problem of structural instability, originally illustrated by Lucas. In addition, our microeconomic data set allows us to pay careful attention to the different ways in which expectations affect wages. We are able to specify the exact horizon of the relevant expectation, to account for the effects of errors in past inflationary expectations, and to match the date when the expectation was formed with the date when the contract was agreed upon.

Our results show that the commercial forecast data can be reliably correlated with union wages in a manner consistent with economic theory. The equations using these direct measures of expectations do perform better than the same specification where lagged inflation is used as an indirect measure of expected inflation and an alternative time-series model of adaptive expectations. The latter specifications, however, perform remarkably well. This is particularly surprising because the commercial forecasts of inflation are based on different information than just lagged inflation, and the correlation between lagged inflation and the forecast data is not stable over our sample period. Our data therefore do not provide strong support to allow us to reject the hypothesis that inflationary expectations are backward looking.

In addition to the effect of inflationary expectations on union wages, we find evidence that union wages respond to changes in labor market conditions which are expected to occur during the life of the contract. When this variable is included, we also find a significant and important effect of current labor market conditions on long-term, escalated contracts. We believe that previous researchers using U.S. data (Daniel Mitchell, 1978, and Wayne Vroman, 1981) have failed to find this latter effect because they have not included the effects of expected changes in labor market conditions.

While our results do not allow us to reject the pessimistic conclusions of George Perry (1978), Mitchell (1978), and Stephen McNees (1979) that inflationary expectations are backward looking, our results do provide some support for proponents of the "credibility effect" (see Fellner) who argue that the perception of both present and *future* demand management policies (and thus future labor market conditions) affect current inflation.

Our analysis is also of interest because we have modeled the effects of cost-of-living adjustments (*COLAs*) differently and in more detail than previous researchers. Our specifi-

*Assistant Professor of Economics, Smith College, Northampton, MA 01063, and Associate Professor of Economics, Amherst College, Amherst, MA 01002. We thank Ralph Beals, Yolanda Henderson, Walter Nicholson, and Beth Yarbrough for comments; any remaining errors are our responsibility.

cation allows us to regress a measure of expected *ex ante* wage change over the contract period solely on variables which are known to agents at the time of the contract negotiations. Previous research on *COLA* contracts has suffered, in our view, by regressing actual, *ex post* wage changes on variables which are unknown to agents at the time of negotiation, thereby introducing spurious correlation.

In Section I, we briefly review the problems of estimating models in the presence of rational expectations and explain how our analysis seeks to overcome these problems. In Section II, we describe our model of union wage determination. The results of previous research on union wage determination are reviewed in the third section, and in the final section we describe the data and analyze our results.

I. Problems in Estimating Models with Rational Expectations

The problems associated with macroeconometric models containing expectations were described in a famous article by Lucas. That article has stimulated a substantial literature analyzing alternative procedures for overcoming these problems. In a companion piece to this paper (see our 1983 article), we have critically reviewed the proposed alternative procedures. In this section we shall briefly discuss the nature of these problems and explain how the use of direct measures of expectations with microeconomic data can avoid them. While we focus our discussion on measures of inflationary expectations, the same points are relevant to measures of expected future labor market conditions.

The two interrelated problems raised by macroeconometric models using indirect measures of expectations are 1) distinguishing between expectational effects and other effects that give rise to distributed lags, and 2) identifying relationships which are structural in the sense of being invariant to changes in policy regimes. Both problems result from the use of indirect measures of expectations. The classic example of an indirect measure of expected inflation is the adaptive expectations model where expected inflation is assumed to be an invariant weighted average of past inflation. Adaptive expectations are used to justify distributed lags in many macroeconomic relationships, such as distributed lags on disposable income in the consumption function, and distributed lags on the average product and user cost of capital in investment functions.

The adaptive expectations model is not the only example of an indirect measure of expectations. For example, Michael Wachter and Susan Wachter (1978), Robert Barro (1977), and many others have argued that expected inflation can be measured by lagged rates of money growth. These measures are also indirect because they, like those in the adaptive model, explicitly define a mechanism by which past data are used to obtain expectations of the future. As will be discussed below, it is important to note that the use of indirect measures also implicitly assumes that the expectation-generating mechanism is stable throughout the sample period, that is, the same past data are used in the same way.

The first problem of macroeconometric models using indirect measures of expectations arises because there are other, nonexpectational reasons for the presence of distributed lags in macroeconomic relationships. For example, distributed lags in investment equations are justified on the basis of technological factors such as costs of rapid adjustment; distributed lags of past inflation in aggregate wage equations are frequently justified on the basis of multiperiod wage contracts. We shall refer to the nonexpectational reasons for distributed lags as inertial effects. Only a small fraction of the macroeconometric literature of the 1960's and early 1970's was concerned with distinguishing between inertial and expectational effects (see, for example, Robert Gordon, 1971).

If one believes that the adaptive model is structural, then the distinction between expectational and inertial effects is not of any particular importance. For simplicity we use the term backward-looking expectations to stand for this view. Backward-looking expectations, however, are not in general consistent with rational expectations. Lucas's contribution was to show how crucial it is to make the distinction between expectational and inertial effects if one assumes expectations are rational. Lucas showed that a macroeconometric model that fails to draw this distinction will suffer from the second problem and thus will be useless for policy purposes. If expectations are rational, the distributed lags that are measuring expectational effects will not be invariant to actual or contemplated changes in policy regimes.

Although Lucas argued that the instability that results from contemplated changes in policy regimes made existing macroeconometric models useless for policy simulations, his point is equally important for estimating macroeconometric models with historical data. One cannot develop a hypothesis of a stable macroeconometric model using indirect measures of expectations over a particular historical sample period unless one is willing to assume that policy has also been stable over the sample period.

Much of the literature that has evolved from the Lucas article has ignored the significance of Lucas's point for estimating models over any historical period. The literature that has developed has proposed two basic alternative approaches to traditional macroeconometrics. One approach (Christopher Sims, 1980; Lucas and Thomas Sargent, 1978) eschews the use of economic theory in developing estimating equations; the other approach (for example, Kenneth Wallis, 1980; Sargent, 1981) uses the implications of rationality to distinguish between expectational and inertial effects. As we argue in our earlier article, both of these alternatives continue to measure expectations indirectly. Thus, statistical models using these approaches require the assumption that policy has followed a stable regime throughout the sample period. This assumption strikes us as particularly unfortunate given recent economic events.

The problems raised by Lucas disappear when one uses direct measures of expectations. A direct measure either is a forecast or directly reflects a forecast. These forecasts are based on past information, but the way in which past information is used may vary over time. Survey data on expectations and commercial forecasts are obvious examples of direct measures. Less obvious examples are nominal interest rates as direct measures of expected inflation, and Tobin's "q" as a direct measure of the expected, future productivity of capital. With a direct measure of expectations one need not infer how past data are used to generate the expectation, nor to assume that the expectation-generating mechanism is stable over time. Thus the problem of distinguishing inertial effects from expectation effects is solved because with direct measures of expectations, expectational effects do *not* give rise to distributed lags. The problem of structural instability is solved because with direct measures of expectations the stability of the expectation-generating mechanism is irrelevant.

The problem with direct measures of expectations is the limited nature of the data. We do not have data on the expectations of inflation for particular firms and groups of workers. One must assume that individual expectations are reliably related to some available direct measure. Therefore, the question remains whether there exist direct measures of expectations that can be used to explain important economic behavior. This is obviously an important question and one which this paper seeks to answer.

Before we leave this section we should point out that goodness of fit should not be the only criterion for judging which measures of expectations can be reliably used to explain economic behavior. For example, Sargent (1976) has shown that for a particular policy regime some adaptive expectations models can be observationally equivalent to rational expectations models. If we are right, however, that recent policy has not followed a stable regime, then our direct measures should have an advantage. In addition, however, the use of direct measures should be capable of explaining within-sample variation with a structure that is stable over time.

II. Union Wage Determination with *COLA*s

We shall follow the literature on union wage determination in separating the determinants of wages into two categories: the real factors which affect real wages; and past and expected future rates of inflation. Before one can specify the latter set of determinants, however, one must pay some attention to the definition of negotiated wages in the presence of cost-of-living adjustments. The previous literature (see particularly Vroman and Mitchell, 1978) has either excluded *COLA* adjustments or, more frequently, included actual, subsequent *COLA* payments. The former method is unsatisfactory because it excludes in many cases the largest part of negotiated wage changes. The latter choice is unsatisfactory because the dependent variable is then determined by factors (i.e., inflation), subsequent to the wage negotiations. Including actual *COLA* payments is particularly troublesome because most investigators also include subsequent inflation rates as independent variables. If one is interested in explaining how actual union wages are determined, this procedure may be appropriate. If, on the other hand, one is interested in explaining how union wage changes are negotiated, this procedure introduces a spurious correlation.

In this paper, we are interested in how union wages are negotiated and the role expectations play in the negotiations. Therefore, neither of the widely used alternatives is appropriate. One can, however, infer from the language of the *COLA* agreement the percentage increase in wages, θ, that results from a one-percentage-point increase in inflation. We argue that rational unions and firms know θ and understand that the expected wage change under the contract is the non-*COLA* portion plus θ times expected inflation.[1] Expected wage change is therefore the dependent variable we wish to explain, and our basic equation can be written

$$(1)\qquad \dot{w} + \theta\pi^e = f(\cdot) + g(\cdot) + v,$$

where $\dot{w}$ = the non-*COLA* change in wages expressed as an annualized percentage of previous wages; π^e = expected inflation over the life of the contract expressed at an annualized rate; $f(\cdot)$ is a function of real factors observed at or before the time the contract was negotiated; $g(\cdot)$ is a function of past and expected future rates of inflation, which are observed at or before the time the contract was negotiated; and v is a residual.

We follow the previous literature (for example, Gordon Sparks and David Wilton, 1971, and W. Craig Riddell, 1979) in specifying the arguments of $f(\cdot)$ with one notable exception:

$$(2)\qquad f(\cdot) = a_1 + \frac{a_2}{U^e} + a_3\left(\frac{U^e_F - U^e}{L}\right),$$

where U^e is the expectation of the current quarter's unemployment rate, U^e_F is the current expectation of the unemployment rate in the final quarter of the current contract, and L is the length of the contract, in years.

The one significant exception is our inclusion of the expected annual change in labor market conditions over the life of the contract where previous researchers have used past changes in the unemployment rate. Since some of our contracts exceed three years in length, it seems important to include more than just current and past labor market conditions. Our forecast data allow us to do this.

The specification of the effects of inflation is given by

$$(3)\qquad g(\cdot) = a_4\pi^e + a_5(1-a_4)\pi^e_{-L} + a_5(\pi_{-L} - \pi^e_{-L}) - \theta_{-1}(\pi_{-L} - \pi^e_{-L}),$$

or

$$= a_4\pi^e + a_5(\pi_{-L} - a_4\pi^e_{-L}) - \theta_{-1}(\pi_{-L} - \pi^e_{-L}),$$

where π_{-L} = the actual inflation over the life of the previous contract, at an annual rate; π^e_{-L} = the rate of inflation which was expected to occur over the life of the previous contract at an annual rate, and θ_{-1} = the *COLA* protection factor θ for the previous contract.

The specification of equation (3) is based on assumptions about the role and effect of the *COLA* clause. We assume that the *ultimate* compensation for inflation does not depend on the size of the *COLA* protection factor θ. The *COLA* clause is designed to index the contract partially for the effects of

[1] In Section IV, we describe the way in which we calculated θ for both capped and uncapped *COLA* clauses.

actual inflation (π). However, we also assume that unions and firms are aware of the effects of the *COLA* clause. They therefore negotiate an appropriate non-*COLA* change in wages ($\dot{w}$), given the *ex ante* expected value of the change in wages provided by the *COLA* ($\theta\pi^e$). As a result, we are assuming that the extent to which the *ex ante*, expected change in wages compensates for expected inflation (viz, the size of a_4) is independent of the size of the *COLA* protection factor θ. Notice this specification does *not* require that firms with different θs negotiate the same level of real wages. Rather, it requires that given the levels of real wages, firms facing the same expected inflation will negotiate the same expected *percentage change* in wages (viz, $\dot{w} + \theta\pi_e$).

The last three terms in the first version of (3) are catch-up terms. The last two of these terms capture the effect on current wages of the unexpected inflation that occurred during the previous contract. In virtually all contracts, the *COLA* protection factor θ is less than one. Thus when π_{-L} exceeds π^e_{-L}, actual real wages will fall. Note, however, that since the *COLA* clause does depend on actual inflation, workers will have automatically received some compensation ($\theta_{-1}(\pi_{-L} - \pi^e_{-L})$) for the unexpected inflation during the previous contract. Current wages, therefore, will only be adjusted for the past unexpected inflation for which compensation was not automatically provided for through the *COLA* clause.[2]

The remaining catch-up term on the first line of (3) allows for the possibility that, *ex ante*, expected wages may not adjust fully to expected inflation. Given the possibility of non-pure inflation, there may only be partial *ex ante* adjustment for expected inflation with further adjustments being made after the actual inflation rate has been observed. It is easiest to see this effect by assuming a_4 is less than one and considering the case where $\pi^e_{-L} = \pi_{-L}$. Then the catch-up becomes $a_5(\pi_{-L} - a_4\pi_{-L})$.

Note that if one knew that all inflations would be pure inflations, then neoclassical theory would predict that $a_4 = a_5 = 1$. In this case the catch-up term becomes $(1 - \theta_{-1})(\pi_{-L} - \pi^e_{-L})$. Current wages would only be adjusted for past inflation if that inflation had been unexpected. With the possibility of non-pure inflations, neither a_4 nor a_5 need equal one. We chose this specification because it implies that for any value of a_4 or a_5 the ultimate compensation for inflation does not depend on the value of θ.[3] We test whether our data are consistent with this specification when we estimate our equation in Section IV.

Combining equations (2) and (3) yields the nonlinear equation which we estimated:

$$\text{(4)}\quad \begin{aligned} \dot{w} + \theta\pi^e = {} & a_1 + a_2/U^e + a_3\big((U^e_F - U^e)/L\big) \\ & + a_4\pi^e + a_5\big(\pi_{-L} - a_4\pi^e_{-L}\big) \\ & - \theta_{-1}\big(\pi_{-L} - \pi^e_{-L}\big) + v. \end{aligned}$$

III. Summary of Previous Research

We are not the first to use either direct measures of expectations or the individual

[2] If contract lengths are constant throughout time, the specification of equation (3) is unambiguous. In those cases in which the negotiated length of the new contract is different than that for the preceding contract, we assumed that negotiators try to catch up for the total (nonannualized) amount of the last contract's unexpected inflation during the life of the current contract.

[3] Those readers familiar with the literature on wage equations might be tempted to specify equation (3) as

$$g(\cdot) = a_4\pi^e + a_5[\pi_{-L} - a_4\pi^e_{-L} - \theta_{-1}(\pi_{-L} - \pi^e_{-L})],$$

where the term in brackets represents the inflation in the previous contract period for which workers were not compensated. If $a_5 \neq 1$, however, this specification implies that the fraction of last period's unexpected inflation for which workers are ultimately compensated varies with the size of θ. It is therefore inappropriate to include the last term of equation (3) within the brackets of the second term. To illustrate this point, consider the two extreme cases where $\theta = 0$ (no *COLA* protection) and $\theta = 1$ (full *COLA* protection). In the first case, the catch-up term becomes $a_5(\pi_{-L} - a_4\pi^e_{-L}) = a_5[(\pi^e_{-L} - a_4\pi^e_{-L}) + (\pi_{-L} - \pi^e_{-L})]$. If $\theta = 1$, on the other hand, workers are automatically and fully compensated for unexpected inflation during the preceding contract period by the amount $\pi_{-L} - \pi^e_{-L}$ and the catch-up term becomes $a_5[\pi^e_{-L} - a_4\pi^e_{-L}]$. If $a_5 \neq 1$, the ultimate compensation for the two contracts will therefore vary with the size of θ.

contract data presented in the U.S. Department of Labor's *Wage Chronologies*. In this section we summarize some of the previous research and describe the ways in which our model and estimations differ from previous work. Although economists have estimated aggregate wage equations using direct measures of expectations (for example, Stephen Turnovsky and Wachter, 1972), we will concentrate on those studies which have used data on individual wage contracts.

The earliest attempts to use individual contract data in wage equations were made by Daniel Hamermesh (1970, 1972), Albert Schwenk (1971), and Sparks and Wilton. Mitchell (1978, 1980) conducted a more recent study in this tradition. The model specifications used by these investigators were essentially backward looking and included a variety of macroeconomic and firm- and industry-specific variables. To explain the *COLA*-augmented wage changes in his later paper, Hamermesh included the contemporaneous changes in prices as an independent variable. This began a tradition of using *ex post* inflation rates to explain *ex post* wage changes. Because actual inflation *determines* the *COLA*-augmented wage changes and both are unknown at the time the contract is negotiated, this specification is inadequate to describe wage negotiations.[4] These studies generally found that past inflation is an important determinant of union wages. Current labor market conditions were occasionally significant determinants of wages in short-term contracts, but not for those determined by long-term contracts.

Two recent studies have included direct measures of inflationary expectations in wage equations using contract data. Riddell analyzed 2,360 Canadian contracts over the period 1953–73. His direct measures of inflationary expectations were the one-year U.S. inflation forecasts from the Livingston survey, as modified by John Carlson (1977).[5] Vroman also used the Livingston data to analyze wages for 2,274 U.S. contracts negotiated between 1957 and 1979. Vroman used many contracts for which the absolute wage increase was known but the base wage had to be estimated. He computed his estimates for the base wage from industry wage data and estimates of union-nonunion differentials.

Vroman and Riddell's specifications of the wage equation were similar. The annualized percentage wage change, including the *ex post COLA* adjustment, was regressed against the inverse of the unemployment rate, the Livingston inflation forecast, and unexpected inflation during the previous contract. Riddell also included past changes in unemployment and a measure of inflation uncertainty, and Vroman included the *ex post* unexpected inflation in the current contract and industry profit rates.

Both researchers find that expected inflation is an important determinant of union wages. In fact, in Riddell's equation the coefficient on expected inflation over the entire sample period is significantly greater than one. Riddell's study is unique in that he finds a large, significant effect of current labor market conditions for all contracts. His study is also interesting because he tests his equations for structural stability over time, and his data allow him to reject the hypothesis of a stable structure.[6] Vroman's study is interesting because he allows the coefficient on his *ex post* unexpected inflation term to differ depending upon whether the contract has a

[4] Although Mitchell briefly experimented with some proxies for expected inflation, such as lagged money growth, he does not deal extensively with the effects of expectations.

[5] In fn. 28 of his article Riddell notes that he used two additional variables for expected inflation in his dissertation. One of these, the Michigan survey series, also represents inflationary expectations in the United States. The other is apparently an ARIMA model using Canadian data and incorporating rational expectations.

[6] In Riddell's early and late subsamples the coefficient on expected inflation generally becomes insignificantly different from one, and the coefficient on unemployment declines by 50 percent or more. In this latter respect, his subsample results are closer to those obtained by D.A.L. Auld et al. (1979; 1981) who used essentially the same Canadian contract data as Riddell and an adaptive measure of expectations. Riddell also estimated backward-looking equations which used lagged price changes as proxies for expected inflation. Although the equations with the Livingston index had much lower standard errors, Riddell did not test for the temporal stability of the backward-looking model.

COLA clause and whether the *COLA* clause is capped. He finds significant differences among these coefficients.

Our study differs from these previous studies because: 1) our dependent variable is the *ex ante* expected wage change; 2) we have a more detailed description of how the terms of the *COLA* clause (viz, the value of θ) affect union wages; 3) we are better able to match the date when the forecast is made with the date when the contract is negotiated; 4) our direct forecasts extend eight rather than four quarters into the future; and 5) we include the effects of expected changes in labor market conditions. Our more detailed specifications, however, require data that limit the number of our observations. These data are described in the next section.

IV. Data and Results

The specification of our wage equation requires all of the following: disaggregated information on the duration of individual contracts; separate data on wage increases resulting in general wage increases and from *COLA*s; and direct measures of expected inflation and labor market conditions.

For our study we used the U.S. data which are contained in the series of *Wage Chronologies* compiled by the Bureau of Labor Statistics. Until the series was terminated last year, the BLS maintained 32 chronologies, each of which presented a detailed history of the negotiations involving one or more major union and employer. Some of these data were given to us by Daniel Mitchell, but other time-series data were constructed for both general and *COLA* wage increases within each contract. Where possible we chose an occupation in the middle of the wage structure for the reference or base wage. Using these contract data it was possible to pinpoint the month in which each contract was negotiated and the month in which it became effective.

A list of the 27 contract pairs which was used is given in Table 1. In those cases in which the *Wage Chronology* for a particular contract pair had not been updated, we obtained recent data from various issues of *Current Wage Developments*. As noted by previous researchers (Hamermesh, 1970, and Mitchell, 1980), reasonable data series cannot be constructed for some of the wage chronologies.

Table 1—Contract Pairs Contained in Sample

Contract pairs having at least one contract in sample with automatic cost-of-living adjustments:

Western Union Telegraph and the Telegraph Workers
Boeing and the International Association of Machinists
F.M.C. and the T.W.U.A.
Western Greyhound Lines and the Amalgamated Association of Street, Electric Railway, and Motor Coach Employees (in recent years the ATU)
Alcoa and the United Aluminum Workers and United Steelworkers
Firestone and the United Rubber Workers
Ford and the United Auto Workers
Pacific Gas and Electric and the International Brotherhood of Electrical Workers
Bituminous Coal Mine Operators and the United Mine Workers
Bethlehem Steel and the Industrial Union of Marine and Shipbuilding Workers
Martin Marietta and the United Automobile, Aerospace, and Agricultural Implement Workers
Lockheed-California and the International Association of Machinists and Aerospace Workers
Anaconda and the United Steelworkers
Pacific Coast Shipbuilders and various unions
U.S. Steel and the United Steelworkers
International Harvester and the United Automobile, Aerospace, and Agricultural Implement Workers
Rockwell International and the United Automobile, Aerospace, and Agricultural Implement Workers
Armour and the Amalgamated Meat Cutters and Butcher Workmen
Class I Railroads and the National Boilermakers and Blacksmiths Union

Other contract pairs in sample:

Pacific Maritime Association and the International Longshoremen's and Warehousemen's Union
Berkshire Hathaway and the Clothing and Textile Workers
International Paper and the Paperworkers and Electrical Workers
Commonwealth Edison and the International Brotherhood of Electrical Workers
Atlantic Richfield and the Oil Workers
Massachusetts Shoe Manufacturers and United Shoe Workers of America
Dan River and the Textile Workers
North Atlantic Shipping and the International Longshoremen's and Warehousemen's Union

We also calculated a *COLA* protection factor θ for each contract. Because *COLA*s often do not begin until the second year of the contract and are not granted in every quarter thereafter, calculations of θ based solely on the negotiated cents per hour wage increase per point increase in the Consumer Price Index are inaccurate. We calculated θ for those contracts whose *COLA*s were not capped by dividing the actual percentage increases resulting from the *COLA* by the actual percentage increase in prices during the contract period. Strictly speaking, our estimates of θ are based on *ex post* information. We assume that rational workers and rational firms will understand the implications of the *COLA* clause for the degree of inflation protection. Note, however, that our dependent variable is expected wage increase, or $\dot{w}+\theta\pi^e$, rather than the actual *ex post* wage increase $\dot{w}+\theta\pi$ used by previous researchers. Using the latter variable as the dependent variable requires assuming not only that workers and firms understand the implications of the *COLA* clause, but that they also have perfect foresight concerning the subsequent inflation rate.

Several of the contracts had cost-of-living adjustments with caps which were triggered at rates of inflation below the rate which was expected to occur during the contract period. In each of these cases the cap was eventually triggered. Since unexpected inflation during the current contract period was uncompensated in these contracts, we set the θs equal to zero and included the capped *COLA* increase in the general wage increase $\dot{w}$.

For our measures of expected inflation and expected labor market conditions we used the Consumer Price Index (*CPI*) and unemployment rate forecasts produced by one of the prominent commercial forecasters. Unlike the Livingston and other survey data, these forecasts are sold to subscribers and therefore pass a market test. The fact that unions and profit-maximizing firms purchase these forecasts is strong a priori evidence that they contain valuable information.[7] The United Auto Workers, for example, negotiates over 2,000 contracts. Lee Price, an economist with the U.A.W., told us in a telephone conversation that the union subscribes to the Blue Chip Indicators collection of forecasts. These estimates are then actually used in bargaining.

There are other desirable attributes of the commercial forecasts. First, McNees, in collecting the data, has taken considerable care to keep track of the date when the data were released. The company alters its forecasts twice each quarter. Thus, for any month in which a contract was negotiated we were able to use the most recent forecast with confidence that we were not including future information. Second, the commercial data also provide information about expected labor market conditions. Finally, as we noted in Section II, multiyear contracts require information on expectations extending more than one year into the future. For each variable we have eight forecasts extending one to eight quarters into the future. While this is an improvement over one-year forecasts, many of our contracts involve expectations extending beyond eight quarters. For these contracts we assumed that the unemployment rate is expected to remain at the level of the eight-quarter forecast. To compute the expected rates of inflation we extended the quarterly rate of inflation implicit in the eight-quarter forecast.[8]

Unfortunately, the commercial forecasts of unemployment and inflation began in 1970–71. Because our model requires a measure of inflationary expectations in the preceding contract period, we therefore had to limit our sample to those 83 contracts for which expectations data were available for the *preceding* contract.[9] The negotiation dates for the 83 contracts extended between June 1972 and December 1981. The sample includes 20

[7] Even though the models used to generate these forecasts are often based on equations which utilize adaptive expectations, the final forecasts are adjusted judgmentally and are therefore direct measures.

[8] Our dependent variable, the expected annual percentage increase in wages during the current contract, uses the wage prior to the newly negotiated increase as its base. Past rates of inflation, however, use previous consumer price levels as their bases. For the independent variables which determine the wage catch-up, we therefore recalculated the preceding rates of actual and expected inflation using the current *CPI* level as the base.

[9] We omitted two observations because they involved contracts whose durations were 6 months or less.

Table 2—Basic Equation using Direct Expectations

	Contracts		
	All	*COLA*	Non-*COLA*
Constant	−.057	−.090	.065
	(.027)	(.035)	(.044)
$1/U^e$	.470	.719	−.030
	(.154)	(.210)	(.192)
$(U_F^e - U^e)/L$	−.035	−.037	−.005
	(.008)	(.017)	(.009)
π^e	1.120	.887	.399
	(.127)	(.182)	(.215)
$(\pi_{-L} - a_4\pi^e_{-L})$	.342	.942	−.001
	(.108)	(.239)	(.111)
SSR	.0410	.0146	.0100
R^2	.468	.673	.148
Number of Observations	83	44	39
			$F(5, 73) = 9.75$[a]

Notes: Standard errors are shown in parentheses. The dependent variable is $\dot{w} + \theta\pi^e + \theta_{-1}(\pi_{-L} - \pi^e_{-L})$.

[a] F-statistic for test of the hypothesis that all coefficients are the same for *COLA* and non-*COLA* contracts.

contracts with durations ranging from 9 to 15 months, 10 two-year contracts, and 53 contracts ranging in length from 30 to 44 months. Forty-four of the contracts had cost-of-living adjustment clauses. These 44 contracts ranged from 24 to 44 months in duration and were negotiated between June of 1973 and December of 1981.

The results of estimating equation (4) with the data from all 83 contracts are given in Table 2.[10] The results in the first column are consistent with our a priori hypotheses. The results in the second and third columns, however, indicate that our wage equation cannot explain the non-*COLA* contracts. We suspect, however, that the non-*COLA* contracts in our sample are not representative of all non-*COLA* contracts.[11]

On the basis of the results in Table 2 we conducted our tests of different measures of expectations with only the *COLA* contracts. The results of these tests are given in Tables 3 and 4. In Table 3 we used the commercial forecasts of both inflation and unemployment as independent variables. In the full sample of *COLA* contracts, all the coefficients have the expected signs and are statistically significant. We jointly tested the two hypotheses that the coefficient on expected inflation during the preceding contract is equal to $-a_5a_4$ and the coefficient on $\theta_{-1}(\pi_{-L} - \pi^e_{-L})$ is minus one, and these hypotheses cannot be rejected at the 5 percent level. Since our dependent variable includes π^e as one of its arguments, it is not surprising to find a significant coefficient on expected inflation. It should be noted, however, that the average θ in our sample of *COLA* contracts is .66 and the range is 0 to 1.0.[12] The fact that the estimated coefficient on expected inflation is so close to (and insignificantly different from) 1.0 is strong evidence that workers do not suffer from money illusion in the current contract period. The coefficient on inflationary surprises during the previous contract period is higher than that obtained by Vroman, and is not significantly different from 1.0.

The coefficients on the current and expected changes in unemployment reveal an important effect on wages. Together they imply that a one-percentage-point increase in the official unemployment rate from 6 percent that was expected to persist over the life

[10] The equations in Tables 2–6 were estimated using nonlinear least squares. We did not use a generalized least squares pooled cross-section time-series method because the number of observations in the sample for each contract pair listed in Table 1 rarely exceeded two.

[11] We thought that part of the problem might be due to the fact that many of our non-*COLA* contracts are in declining industries. We included an industry profits variable to try to account for the special nature of these contracts. Although this variable was significant in the non-*COLA* equation, all the other variables were insignificant. When the profits variable was included in the basic equation for the *COLA* contracts, however, it was insignificant. The results of these regressions are essentially the same as those reported in Tables 3 and 5 with one exception, which we discuss in fn. 15. Although previous researchers have employed different specifications, it is still surprising that they found significant effects of inflation and unemployment in non-*COLA* contracts. Because we used direct measures of expectations our sample period was limited to the 1970's, a decade in which the percentage of workers under major contracts who were not protected by *COLA*s fell by about half. As a result, non-*COLA* contracts in the 1970's represent a significantly different population than non-*COLA* contracts in the previous decade.

[12] As described in the text, we set $\theta = 0$ for those capped contracts in which the cap was both triggered at rates of inflation below π^e and actually triggered during the contract period.

TABLE 3—CONTRACTS WITH *COLA*s
(Direct Expectations)

	Contracts			
	All *COLA*	All *COLA*	Early *COLA*	Late *COLA*
Constant	−.090	−.040	−.088	.007
	(.035)	(.030)	(.059)	(.114)
$1/U^e$	.719	.435	.714	.059
	(.210)	(.173)	(.320)	(.759)
$(U_F^e - U^e)/L$	−.037	–	−.029	−.009
	(.017)		(.037)	(.035)
π^e	.887	.684	.760	1.009
	(.182)	(.125)	(.376)	(.322)
$\pi_{-L} - a_4\pi^e_{-L}$	.942	.982	1.100	.585
	(.239)	(.222)	(.600)	(.466)
SSR	.0146	.0164	.0063	.0076
R^2	.673	.632	.553	.744
Number of Observations	44	44	22	22
Sample Period	6/73–12/81	6/73–12/81	6/73–11/77	11/77–12/81
	$F(2,37) = .24$[a]	$F(2,38) = 1.59$[a]	$F(2,15) = 4.15$[a]	$F(2,15) = 9.6$[a]
			$F(5,34) = .30$[b]	

[a] *F*-statistic for test that the coefficient on $\pi^e_{-L} = -a_4a_5$ and the coefficient on $\theta_{-1}(\pi_{-L} - \pi^e_{-L}) = -1$.

[b] *F*-statistic for test of stability of all coefficients over time.

of the contract would slow the annual increase in wages by 1.7 percentage points. If, however, unemployment was expected to return to 6 percent before the new contract expired, there would only be a .5 percentage point deceleration in annual wage increases for a three-year contract.

We believe that previous researchers using U.S. data have failed to find significant unemployment effects in long-term contracts because they have not included variables measuring expected future changes in unemployment. If unemployment is currently high or low relative to trend, theory predicts it will move toward its trend level in the future. Since the omitted variable is likely to be negatively correlated with current unemployment, previously estimated coefficients on unemployment are likely to be biased toward zero. To test this hypothesis we reestimated our basic equation over all *COLA* contracts but excluded the expected change in unemployment as an independent variable. In the second column of Table 3, where the results are presented, note that the coefficient on the inverse of the current unemployment rate falls by 40 percent. The problem of omitting the expected change in unemployment should be less severe the shorter the contract. Mitchell (1978) found that a permanent increase in unemployment from 6 to 7 percent would decrease wages in *short-term contracts* by 1.2 percentage points per year, a figure which is close to our estimate for all *COLA* contracts of a permanent increase in the unemployment rate.

In the third and fourth columns of Table 3 we test our model for structural stability by reestimating equation (4) over the earlier and later halves of the sample. A few of the coefficients exhibit instability, but the overall *F*-statistic is well below the 5 percent critical value of 2.5. Note, however, that the coefficient on the catch-up term $\pi_{-L} - a_4\pi^e_{-L}$ declines sharply in the second half of the sample. This may indicate an inability of workers to catch up for unanticipated inflation during this period, even if the inflation is pure inflation. On the other hand, it may represent an increase in non-pure inflation during this period and merely reflect workers' inability to keep real wages constant during such periods.

In order to perform a simple test of this latter hypothesis, we reestimated our model using percentage changes in the *CPI* for all commodities and services excluding energy for lagged actual inflation in the catch-up

TABLE 4—CONTRACTS WITH *COLA*s
(Direct Expectations: Lagged Inflation Excluding Energy)

	Contracts		
	All *COLA*	Early *COLA*	Late *COLA*
Constant	−.082	−.087	.0039
	(.034)	(.059)	(.107)
$1/U^e$	.672	.720	.093
	(.203)	(.326)	(.709)
$(U^e_F - U^e)/L$	−.034	−.035	−.009
	(.017)	(.039)	(.034)
π^e	.906	.841	.983
	(.181)	(.434)	(.287)
$\pi_{-L} - a_4\pi^e_{-L}$	1.009	1.049	.712
	(.262)	(.685)	(.475)
SSR	.0144	.00656	.00737
R^2	.677	.536	.753
Number of Observations	44	22	22
Sample Period	6/73–12/81	6/73–11/77	11/77–12/81
	$F(2,37) = .17^a$	$F(2,15) = 3.94^a$	$F(2,15) = 9.40^a$
		$F(5,34) = .21^b$	

a,bSee Table 3.

term $\pi_{-L} - a_4\pi^e_{-L}$. This specification is based on the presumption that most of the non-pure inflation during the 1970's was due to energy price increases.[13] The results are presented in Table 4. The equation for the complete sample performs slightly better than that given in Table 3. The sum of squared residuals is slightly lower, the coefficients on the inflation terms are closer to 1.0, and the *F*-test for the constraints on the two coefficients is very close to zero.

When the sample is split, the last two columns indicate that the coefficients on the labor market variables remain unstable, but those relating to the inflation terms are much more consistent. The *F*-statistic testing the stability of all the coefficients over time has also declined by 30 percent to 0.21. While these results seem to confirm the hypothesis of non-pure inflation, they should be treated cautiously because of the relatively small number of contracts in each of the two sample splits.[14]

Our results indicate that commercial forecasts of both inflation and unemployment are significant determinants of wage negotiating behavior. Following our discussion in Section I it is also important to compare the performance with results based on indirect measures of expectation. In Table 5 our basic equation is estimated with the rate of inflation in the year previous to the contract being used as an adaptive expectations proxy for expected inflation. The lagged inflation rate is typically used as the measure of adaptive expectations in the previous literature

[13] Because almost all cost-of-living adjustments are indexed to the actual inflation rate, we retained the total rate of inflation in the term representing the amount of unexpected inflation for which workers were unexpectedly compensated during the preceding contract, $\theta_{-1}(\pi_{-L} - \pi^e_{-L})$.

[14] Because direct forecasts of expected inflation are included on both sides of our estimated equation, errors in measurement may lead to inconsistent estimates of the coefficients. Therefore, we also estimated our model using an instrumental variables technique. We used the adaptive time-series forecasts of inflation presented in the next section as instruments for those direct forecasts of inflation which appear as independent variables. The estimated equation was

$$\dot{w} + \theta\pi^e + \theta_{-1}(\pi_{-L} - \pi^e_{-L})$$
$$= \underset{(.044)}{-.152} + \underset{(.250)}{1.026}(1/U^e) - \underset{(.016)}{.028}(U^e_F - U^e)/L$$
$$+ \underset{(.201)}{.869}\pi^e + \underset{(.329)}{1.366}(\pi_{-L} - .869\pi^e_{-L}),$$

where the estimated asymptotic standard errors of the coefficients are shown in parentheses. The sum of squared residuals for this equation was .0164 and the R^2 was .631.

TABLE 5—CONTRACTS WITH *COLA*S
(Adaptive Expectations: $\pi^e = \pi_{-1}$)

	Contracts			
	All *COLA*	All *COLA*	Early *COLA*	Late *COLA*
Constant	−.102	−.050	−.121	−.098
	(.037)	(.033)	(.069)	(.135)
$1/U^e$	.784	.494	.891	.642
	(.224)	(.202)	(.344)	(.854)
$(U_F^e - U^e)/L$	−.044	–	−.015	−.038
	(.022)		(.050)	(.046)
π^e	.964	.717	.557	1.235
	(.249)	(.140)	(.322)	(.396)
$\pi_{-L} - a_4\pi^e_{-L}$	.745	.948	1.458	.460
	(.307)	(.271)	(.920)	(.370)
SSR	.0186	.0209	.0076	.0099
R^2	.771	.743	.743	.809
Number of Observations	44	44	22	22
Sample Period	6/73–12/81	6/73–12/81	6/73–11/77	11/77–12/81
	$F(2,37) = 2.02$[a]	$F(2,38) = 1.88$[a]	$F(2,15) = 4.02$[a]	$F(2,15) = 12.89$[a]
		$F(5,34) = .46$[b]		

[a,b]See Table 3.

(see Mitchell, 1978, 1980; and Vroman). Even though this is a crude proxy, the results of the first column of Table 3 and Table 5 are similar. The expected inflation variable has a value closer to the expected theoretical value of one in the latter but also has a larger standard error. The general results, however, are quite similar in the two regressions. Once again, all the coefficients are significant. The second column of Table 5 again illustrates that the omission of the expected change in unemployment will bias the coefficient on the current unemployment rate towards zero. In judging goodness of fit one must recall that the left-hand side variables are not the same. The lower sum-of-squared residuals of the equations in Table 3, however, imply that in terms of explaining the non-*COLA* portion of wage change the direct expectation equations are superior over all sample periods. Surprisingly, however, the adaptive expectations equation has a very low F-statistic for testing the hypothesis of the time stability of the coefficients.[15]

We were quite surprised at how well our crude adaptive measure had done. While many past researchers have had success using one-year lagged inflation rates as measures of expected inflation, it still surprised us that this measure could perform nearly as well as our carefully constructed and presumably more sophisticated measure. We felt this particularly true given the precise way inflationary expectations had been modeled in equation (4). It seems hard to believe that lagged inflation could be playing the role of a catch-all for expectational and other inertial effects that it might play in aggregate wage equations (see John Taylor, 1980).

Given the relative success of the simple measure of adaptive expectations, we reestimated our equation using a more sophisticated model of adaptive expectations. Ex-

[15]We also tried the adaptive measure on the full sample of *COLA* and non-*COLA* contracts. The regressions with the adaptive measure performed slightly better than those with the direct measure. The non-*COLA* equation, however, was still very bad, even when the profit variable was included. Finally, we ran the *COLA* regressions in Tables 3 and 5 with the profit variable. The results are very similar, except in this case the F-statistic testing for the stability of all the coefficients over time was 1.13 for the equation with the direct measure and .94 for the equation with the adaptive measure.

TABLE 6—CONTRACTS WITH *COLA*s
(Adaptive Expectations: Time-Series Model)

	Contracts			
	All *COLA*	All *COLA*	Early *COLA*	Late *COLA*
Constant	−.115	−.042	−.116	−.137
	(.041)	(.033)	(.077)	(.164)
$1/U^e$	.857	.443	.824	.277
	(.244)	(.180)	(.403)	(.636)
$(U_F^e - U^e)/L$	−.039	–	−.020	−.023
	(.016)		(.031)	(.032)
π^e	.958	.702	.784	2.36
	(.229)	(.181)	(.271)	(1.09)
$\pi_{-L} - a_4\pi^e_{-L}$	1.12	.997	1.66	−.025
	(.271)	(.240)	(.668)	(.486)
SSR	.0147	.0168	.0074	.0060
R^2	.512	.442	.343	.678
Number of Observations	44	44	22	22
Sample Period	6/73–12/81	6/73–12/81	6/73–11/77	11/77–12/81
	$F(2,37) = .59$[a]	$F(2,38) = 2.86$[a]	$F(2,15) = 3.83$[a]	$F(2,15) = 7.65$[a]
		$F(5,34) = .66$[b]		

[a,b] See Table 3.

TABLE 7—CONTRACTS WITH *COLA*s
(Common information in the three measures of expectations)

	Contracts[a]		
	All *COLA*	Early *COLA*	Late *COLA*
Dependent Variable:			
Direct π^e			
Independent Variables:			
Constant	.015	.025	.020
	(.005)	(.002)	(.008)
Adaptive π_e	.674	.489	.678
$(= \pi_{-1})$	(.049)	(.028)	(.070)
SSR	.0036	.0002	.0017
R^2	.819	.940	.823
		$F(2,40) = 17.89$[b]	
Constant	−.041	−.097	−.045
	(.010)	(.024)	(.020)
Adaptive π_e	1.475	2.268	1.507
(ARIMA model)	(.117)	(.325)	(.216)
SSR	.0042	.0010	.0028
R^2	.791	.708	.708
		$F(2,40) = 1.95$[b]	

[a] Number of observations for All *COLA* is 44; for Early *COLA* and Late *COLA*, respectively, 22. The sample period for All *COLA* is 6/73–12/81; for Early *COLA*, 6/73–11/77; for Late *COLA*, 11/77–12/81.

[b] See Table 2.

perimentation showed that the logarithm of monthly levels of the *CPI* between 1968 and 1981 could be reasonably represented by a $(2\ 1\ 0)(0\ 1\ 1)_{12}$ ARIMA model. Using this model we generated forecasts of the price level over the relevant horizon at each contract's date of negotiation, and these forecasts were used to estimate equation (4).

The results of this procedure are given in Table 6. Most of the coefficients and regression statistics are very similar to those in Tables 3 and 5. The sums of squared residuals are only slightly greater than those obtained using direct expectations. Although the F-statistic testing the hypothesis that all the coefficients are stable over time remains well below its critical value, it is more than twice the value obtained using direct expectations. This is especially noteworthy since the time-series model which generated these forecasts uses data for the entire sample period, and thereby includes information which becomes known after the date of negotiation for many contracts in our sample.

Given the similarity in these results, we suspected that our three measures of expectations might contain the same information. We therefore tested how closely correlated our two measures of adaptive expectations were to our direct measure and how stable the correlations were over time. Table 7 presents the results of these tests. While these measures are correlated, the correlation is far from perfect and is unstable over time.

V. Conclusion

We have demonstrated that direct measures of both labor market and inflationary expectations are significant determinants of union wages. The failure to include a measure of the expected change in unemployment over the contract period has also led previous researchers to underestimate the effects of current labor market conditions on wages in long-term *COLA* contracts. Our data, however, do not allow us to reject the hypothesis that inflationary expectations are backward looking. The equations with the direct measures perform better, but the differences are surprisingly small.

REFERENCES

Auld et al., D. A. L., "The Effect of Settlement Stage on Negotiated Wage Settlements in Canada," *Industrial and Labor Relations Review*, January 1981, *34*, 234–44.

______, *The Determinants of Negotiated Wage Settlements in Canada (1966–1975)*, Ottawa, 1979.

Barro, Robert, "Unanticipated Money Growth and Unemployment in the United States," *American Economic Review*, March 1977, *67*, 101–15.

Carlson, John, "A Study of Price Forecasts," *Annals of Economics and Social Measurement*, 1977, *6*, 27–56.

Fellner, William, "The Credibility Effect and Rational Expectations," *Brookings Papers on Economic Activity*, 1:1979, 167–78.

Gordon, Robert J., "Inflation in Recession and Recovery," *Brookings Papers on Economic Activity*, 1:1971, 105–58.

Gramlich, Edward, "Macro Policy Responses to Price Shocks," *Brookings Papers on Economic Activity*, 1:1979, 125–66.

Hamermesh, Daniel, "Wage Bargains, Threshold Effects, and the Phillips Curve," *Quarterly Journal of Economics*, August 1970, *84*, 501–17.

______, "Market Power and Wage Inflation," *Southern Economic Journal*, October 1972, *39*, 204–11.

Kaufman, Roger T. and Woglom, Geoffrey, "Estimating Models with Rational Expectations," *Journal of Money, Credit and Banking*, August 1983, *15*, 275–85.

Lucas, Robert E., "Econometric Policy Evaluations: A Critique," in K. Brunner and A. Meltzer, eds., *The Phillips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl. 1976, 19–46.

______ **and Sargent, Thomas**, "After Keynesian Macroeconomics," in Federal Reserve Bank of Boston Conference Series No. 19, *After the Phillips Curve: Persistence of High Inflation and High Unemployment*, Boston, 1978.

McNees, Stephen, "The Phillips Curve: Forward- or Backward-Looking?," *New England Economic Review*, July/August 1979, 46–54.

Mitchell, Daniel, "Union Wage Determination: Policy Implications and Outlook," *Brookings Papers on Economic Activity*, 3:1978, 537–82.

______, *Unions, Wages, and Inflation*, Washington: The Brookings Institution, 1980.

Perry, George, "Slowing the Wage-Price Spiral: The Macroeconomic View," *Brookings Papers on Economic Activity*, 2:1978, 259–91.

Riddell, William Craig, "The Empirical Foundations of the Phillips Curve: Evidence from Canadian Wage Contract Data," *Econometrica*, January 1979, *47*, 1–24.

Sargent, Thomas, "The Observational Equivalence of Natural and Unnatural Rate Theories of Macroeconomics," *Journal of Political Economy*, June 1976, *84*, 631–40.

______, "Interpreting Economic Time Series," *Journal of Political Economy*, April 1981, *89*, 213–48.

Schwenk, Albert, *The Influence of Selected Industry Characteristics on Negotiated Settlements*, Staff Paper No. 5, Bureau of Labor Statistics, Washington, 1971.

Sims, Christopher, "Macroeconomics and Reality," *Econometrica*, January 1980, *48*, 1–48.

Sparks, Gordon and Wilton, David, "Determinants of Negotiated Wage Increases: An Empirical Analysis," *Econometrica*, September 1971, *39*, 739–50.

Taylor, John, "Aggregate Dynamics and Staggered Contracts," *Journal of Political Economy*, February 1980, *88*, 1–23.

Turnovsky, Stephen and Wachter, Michael, "A Test of the Expectations Hypothesis Using Directly Observed Wage and Price Expectations," *Review of Economics and Statistics*, February 1972, *54*, 47–54.

Vroman, Wayne, "Wage Contract Settlements in U.S. Manufacturing," paper presented at annual meetings of Western Economic Association, San Francisco, July, 1981.

Wachter, Michael and Wachter, Susan, "Institutional Factors in Domestic Inflation," in Federal Reserve Bank of Boston Conference Series No. 19, *After the Phillips Curve: Persistence of High Inflation and High Unemployment*, Boston, 1978.

Wallis, Kenneth, "Econometric Implications of the Rational Expectations Hypothesis," *Econometrica*, January 1980, *48*, 49–74.

U.S. Department of Labor, *Current Wage Developments*, Washington, various issues.

______, *Wage Chronologies*, Washington, various issues.

Equilibrium Unemployment as a Worker Discipline Device

By CARL SHAPIRO AND JOSEPH E. STIGLITZ*

Involuntary unemployment appears to be a persistent feature of many modern labor markets. The presence of such unemployment raises the question of why wages do not fall to clear labor markets. In this paper we show how the information structure of employer-employee relationships, in particular the inability of employers to costlessly observe workers' on-the-job effort, can explain involuntary unemployment[1] as an equilibrium phenomenon. Indeed, we show that imperfect monitoring necessitates unemployment in equilibrium.

The intuition behind our result is simple. Under the conventional competitive paradigm, in which all workers receive the market wage and there is no unemployment, the worst that can happen to a worker who shirks on the job is that he is fired. Since he can immediately be rehired, however, he pays no penalty for his misdemeanor. With imperfect monitoring and full employment, therefore, workers will choose to shirk.

To induce its workers not to shirk, the firm attempts to pay more than the "going wage"; then, if a worker is caught shirking and is fired, he will pay a penalty. If it pays one firm to raise its wage, however, it will pay all firms to raise their wages. When they all raise their wages, the incentive not to shirk again disappears. But as all firms raise their wages, their demand for labor decreases, and unemployment results. With unemployment, even if all firms pay the same wages, a worker has an incentive not to shirk. For, if he is fired, an individual will not immediately obtain another job. The equilibrium unemployment rate must be sufficiently large that it pays workers to work rather than to take the risk of being caught shirking.

The idea that the threat of firing a worker is a method of discipline is not novel. Guillermo Calvo (1981) studied a static model which involves equilibrium unemployment.[2] No previous studies have treated general market equilibrium with dynamics, however, or studied the welfare properties of such unemployment equilibria. One key contribution of this paper is that the punishment associated with being fired is endogenous, as it depends on the equilibrium rate of unemployment. Our analysis thus goes beyond studies of information and incentives within organizations (such as Armen Alchian and Harold Demsetz, 1972, and the more recent and growing literature on worker-firm relations as a principal-agent problem) to inquire about the equilibrium conditions in markets with these informational features.

The paper closest in spirit to ours is Steven Salop (1979) in which firms reduce turnover costs when they raise wages; here the savings from higher wages are on monitoring costs (or, at the same level of monitoring, from increased output due to increased effort). As in the Salop paper, the unemployment in this paper is definitely involuntary, and not of the standard search theory type (Peter Diamond, 1981, for example). Workers have perfect information about all job opportunities in our model, and unemployed workers strictly prefer to work at wages less than the prevailing market wage (rather than to remain unemployed); there are no vacancies.

*Woodrow Wilson School of Public and International Affairs, and Department of Economics, respectively, Princeton University, Princeton, NJ 08540. We thank Peter Diamond, Gene Grossman, Ed Lazear, Steve Salop, and Mike Veall for helpful comments. Financial support from the National Science Foundation is appreciated.

[1] By involuntary unemployment we mean a situation where an unemployed worker is willing to work for less than the wage received by an equally skilled employed worker, yet no job offers are forthcoming.

[2] In his 1979 paper, Calvo surveyed a variety of models of unemployment, including his hierarchical firm model (also with Stanislaw Wellisz, 1979). There are a number of important differences between that work and this paper, including the specification of the monitoring technology.

The theory we develop has several important implications. First, we show that unemployment benefits (and other welfare benefits) increase the equilibrium unemployment rate, but for a reason quite different from that commonly put forth (i.e., that individuals will have insufficient incentives to search for jobs). In our model, the existence of unemployment benefits reduces the "penalty" associated with being fired. Therefore, to induce workers not to shirk, firms must pay higher wages. These higher wages reduce the demand for labor.

Second, the model explains why wages adjust slowly in the face of aggregate shocks. A decrease in the demand for labor will ultimately cause a lower wage and a higher level of unemployment. In the transition, however, the wage decrease will match the growth in the unemployment pool, which may be a sluggish process.

Third, we show that the market equilibrium which emerges is not, in general, Pareto optimal, where we have taken explicitly into account the costs associated with monitoring. There exist, in other words, interventions in the market that make everyone better off. In particular, we show that there are circumstances in which wage subsidies are desirable. There are also circumstances where the government should intervene in the market by supplying unemployment insurance, even if all firms (rationally) do not. A (small) turnover tax is desirable, because high turnover increases the flow of job vacancies, and hence the flow out of the unemployment pool, making the threat of firing less severe.

Additionally, our theory provides predictions about the characteristics of labor markets which cause the natural rate (i.e., equilibrium level) of unemployment to be relatively high: high rates of labor turnover, high monitoring costs, high discount rates for workers, significant possibilities for workers to vary their effort inputs, or high costs to employers (such as broken machinery) from shirking.

Finally, our theory shows how wage distributions (for identical workers) can persist in equilibrium. Firms which find shirking particularly costly will offer higher wages than other firms do. The dual role wages play by allocating labor and providing incentives for employee effort allows wage dispersion to persist.

Although we have focused our analysis on the labor market, it should be clear that a similar analysis could apply to other markets (for example, product or credit markets) as well. This paper can be viewed as an analysis of a simplified general equilibrium model of an economy in which there are important principal-agent (incentive) problems, and in which the equilibrium entails *quantity constraints* (job rationing). As in all such problems, it is important to identify what is observable, and, based on what is observable, what are the set of feasible contractual arrangements between the parties to the contract. Under certain circumstances, for instance, workers might issue performance bonds and this might alleviate the problems with which we are concerned in this paper. In Section III we discuss the role of alternative incentive devices.

In the highly simplified model upon which we focus here, all workers are identical, all firms are identical, and thus, in equilibrium, all pay the same wage. The assumption that all workers are the same is important, because it implies that being fired carries no stigma (the next potential employer knows that the worker is no more immoral than any other worker; he only infers that the firm for which the worker worked must have paid a wage sufficiently low that it paid the worker to shirk). We have made this assumption because we wished to construct the simplest possible model focussing simply on incentive effects, in which adverse selection considerations play no role. In a sequel, we hope to explore the important interactions between the two fundamental information problems of adverse selection and moral hazard.[3]

The assumption that all firms are the same is not critical for the existence of equilibrium unemployment. Firm heterogeneity will, however, lead to a wage distribution. If the

[3] Other studies have focused on quantity constraints (rationing) with adverse-selection problems. See Stiglitz (1976), Charles Wilson (1980), Andrew Weiss (1980), and Stiglitz and Weiss (1981).

damage that a particular firm incurs as a result of a worker not performing up to standard is larger, the firm will have an incentive to pay the worker a higher wage. Similarly, if the cost of monitoring (detecting shirking) for a firm is large, that firm will also pay a higher wage. Thus, even though workers are all identical, workers for different firms will receive different wages. There is considerable evidence that, in fact, different firms do pay different wages to workers who appear to be quite similar (for example, more capital intensive firms pay higher wages). The theory we develop here may provide part of the explanation of this phenomenon.

In Section I, we present the basic model in which workers are risk neutral. Quit rates and monitoring intensities are exogenous. A welfare analysis of the unemployment equilibrium is provided. In Section II, we comment on extensions of the analysis to situations where monitoring intensities and quit rates are endogenous, and where workers are risk averse. Section III compares the role of unemployment as an incentive device with other methods of enforcing discipline on the labor force.

I. The Basic Model

In this section we formulate a simple model which captures the incentive role of unemployment as described above. Extensions and modifications of this basic model are considered in subsequent sections.

A. *Workers*

There are a fixed number, N, of identical workers, all of whom dislike putting forth effort, but enjoy consuming goods. We write an individual's instantaneous utility function as $U(w, e)$, where w is the wage received and e is the level of effort on the job. For simplicity, we shall assume the utility function is separable; initially, we shall also assume that workers are risk neutral. With suitable normalizations, we can therefore rewrite utility as $U = w - e$. Again, for simplicity, we assume that workers can provide either minimal effort ($e = 0$), or some fixed positive level of $e > 0$.[4] When a worker is unemployed, he receives unemployment benefits of $\overline{w}$ (and $e = 0$).

Each worker is in one of two states at any point in time: employed or unemployed. There is a probability b per unit time that a worker will be separated from his job due to relocation, etc., which will be taken as exogenous. Exogenous separations cause a worker to enter the unemployment pool. Workers maximize the expected present discounted value of utility with a discount rate $r > 0$.[5] The model is set in continuous time.

B. *The Effort Decision of a Worker*

The only choice workers make is the selection of an effort level, which is a discrete choice by assumption. If a worker performs at the customary level of effort for his job, that is, if he does not shirk, he receives a wage of w and will retain his job until exogenous factors cause a separation to occur. If he shirks, there is some probability q (discussed below), per unit time, that he will be caught.[6] If he is caught shirking he will be fired,[7] and forced to enter the unemployment pool. The probability per unit time of acquiring a job while in the unemployment pool (which we call the job acquisition rate, an endogenous variable calculated below) determines the expected length of the unemployment spell he must face. While unemployed he receives unemployment compensation of $\overline{w}$ (also discussed below).

[4] Including effort as a continuous variable would not change the qualitative results.

[5] That is, we assume individuals are infinitely lived, and have a pure rate of time preference of r. They maximize

$$W = E\int_0^{\infty} u(w(t), e(t))\exp(-rt)\,dt,$$

where we have implicitly assumed that individuals can neither borrow nor lend. Allowing an exponential death rate would not alter the structure of the model; neither would borrowing in the risk-neutral case.

[6] For now we take q as exogenous; later it will be endogenous. The assumption of a Poisson detection technology, like a number of the other assumptions employed in the analysis, is made to ensure that the model has a simple stationary structure.

[7] This will be firm's optimal policy in equilibrium.

The worker selects an effort level to maximize his discounted utility stream. This involves comparison of the utility from shirking with the utility from not shirking, to which we now turn. We define V_E^S as the expected lifetime utility of an employed shirker, V_E^N as the expected lifetime utility of an employed nonshirker, and V_u as the expected lifetime utility of an unemployed individual. The fundamental asset equation for a shirker is given by

$$rV_E^S = w + (b+q)(V_u - V_E^S), \tag{1}$$

while for a nonshirker, it is

$$rV_E^N = w - e + b(V_u - V_E^N). \tag{2}$$

Each of these equations is of the form "interest rate times asset value equals flow benefits (dividends) plus expected capital gains (or losses)."[8] Equations (1) and (2) can be solved for V_E^S and V_E^N:

$$V_E^S = \frac{w + (b+q)V_u}{r+b+q}; \tag{3}$$

$$V_E^N = \frac{(w-e) + bV_u}{r+b}. \tag{4}$$

The worker will choose not to shirk if and only if $V_E^N \geq V_E^S$. We call this the *no-shirking condition* (*NSC*), which, using (3) and (4), can be written as

$$w \geq rV_u + (r+b+q)e/q \equiv \hat{w}. \tag{5}$$

Alternatively, the *NSC* also takes the form $q(V_E^S - V_u) \geq e$. This highlights the basic implication of the *NSC*: unless there is a penalty associated with being unemployed, everyone will shirk. In other words, if an individual could immediately obtain employment after being fired, $V_u = V_E^S$, and the *NSC* could never be satisfied.

Equation (5) has several natural implications. If the firm pays a sufficiently high wage, then the workers will not shirk. The critical wage, $\hat{w}$, is higher

(a) the higher the required effort (e),

(b) the higher the expected utility associated with being unemployed (V_u),

(c) the lower the probability of being detected shirking (q),

(d) the higher the rate of interest (i.e., the relatively more weight is attached to the short-run gains from shirking (until one is caught) compared to the losses incurred when one is eventually caught),

(e) the higher the exogenous quit rate b (if one is going to have to leave the firm anyway, one might as well cheat on the firm).

C. *Employers*

There are M identical firms, $i = 1,\ldots, M$. Each firm has a production function $Q_i = f(L_i)$, generating an aggregate production function of $Q = F(L)$.[9] Here L_i is firm i's effective labor force; we assume a worker contributes one unit of effective labor if he does not shirk. Otherwise he contributes nothing (this is merely for simplicity). Therefore firms compete in offering wage packages, subject to the constraint that their workers choose not to shirk. We assume that $F'(N) > e$, that is, full employment is efficient.

The monitoring technology (q) is exogenous. Monitoring choices by employers are analyzed in the following section. We assume

[8] A derivation follows: taking V_u as given and looking at a short time interval $[0,t]$ we have

$$V_E = wt + (1-rt)[btV_u + (1-bt)V_E],$$

since there is probability bt of leaving the job during the interval $[0,t]$ and since $e^{-rt} \approx 1 - rt$. Solving for V_E, we have

$$V_E = [wt + (1-rt)btV_u]/[1-(1-rt)(1-bt)].$$

Taking limits as $t \to 0$ gives (1). Equation (2) can be derived similarly.

[9] That is,

$$F(L) \equiv \max_{\{L_i\}} \sum f_i(L_i)$$

such that $\Sigma L_i = L$. This assumes that in market equilibrium, labor is efficiently allocated, as it will be in the basic model of this section. The modifications required for more general cases, when different firms face different critical no-shirking wages, $\hat{w}_i$, or have different technologies, are straightforward.

that other factors (for example, exogenous noise or the absence of employee specific output measures) prevent monitoring of effort via observing output.

A firm's wage package consists of a wage, w, and a level of unemployment benefits, $\bar{w}$.[10] Each firm finds it optimal to fire shirkers, since the only other punishment, a wage reduction, would simply induce the disciplined worker to shirk again.

It is not difficult to establish that all firms offer the smallest unemployment benefits allowed (say, by law).[11] This follows directly from the *NSC*, equation (5). An individual firm has no incentive to set $\bar{w}$ any higher than necessary. An increase in $\bar{w}$ raises V_u and hence requires a higher w to meet the *NSC*. Therefore, increases in $\bar{w}$ cost the firm both directly (higher unemployment benefits) and indirectly (higher wages). Since the firm has no difficulty attracting labor (in equilibrium), it sets $\bar{w}$ as small as possible. Hence we can interpret $\bar{w}$ in what follows as the minimum legal level, which is offered consistently by all firms.

Having offered the minimum allowable $\bar{w}$, an individual firm pays wages sufficient to induce employee effort, that is, $w = \hat{w}$ to meet the *NSC*. The firm's labor demand is given by equating the marginal product of labor to the cost of hiring an additional employee. This cost consists of wages and future unemployment benefits. For $\bar{w} = 0$,[12] the labor demand is given simply by $f'(L_i) = \hat{w}$, with aggregate labor demand of $F'(L) = \hat{w}$.

D. *Market Equilibrium*

We now turn to the determination of the equilibrium wage and employment levels. Let us first indicate heuristically the factors which determine the equilibrium wage level.

If wages are very high, workers will value their jobs for two reasons: (a) the high wages themselves, and (b) the correspondingly low level of employment (due to low demand for labor at high wages) which implies long spells of unemployment in the event of losing one's job. In such a situation employers will find they can reduce wages without tempting workers to shirk.

Conversely, if the wage is quite low, workers will be tempted to shirk for two reasons: (a) low wages imply that working is only moderately preferred to unemployment, and (b) high employment levels (at low wages there is a large demand for labor) imply unemployment spells due to being fired will be brief. In such a situation firms will raise their wages to satisfy the *NSC*.

Equilibrium occurs when each firm, taking as given the wages and employment levels at other firms, finds it optimal to offer the going wage rather than a different wage. The key market variable which determines individual firm behavior is V_u, the expected utility of an unemployed worker. We turn now to the calculation of the equilibrium V_u.[13]

The asset equation for V_u, analogous to (1) and (2), is given by

$$(6) \qquad rV_u = \bar{w} + a(V_E - V_u),$$

where a is the job acquisition rate and V_E is the expected utility of an employed worker (which equals V_E^N in equilibrium). We can now solve (4) and (6) simultaneously for V_E and V_u to yield

$$(7) \qquad rV_E = \frac{(w-e)(a+r)+\bar{w}b}{a+b+r};$$

$$(8) \qquad rV_u = \frac{(w-e)a+\bar{w}(b+r)}{a+b+r}.$$

[10] More complex employment contracts, for example, wages rising with seniority, are discussed in Section III. With our assumptions of stationarity and identical workers, employers cannot improve on the simple employment provisions considered here.

[11] We are implicitly assuming that the firm cannot offer $\bar{w}$ only to workers who quit. This is so because the firm can always fire a worker who wishes to quit, and it would be optimal for the firm to do so.

[12] For $\bar{w} > 0$ the expected cost of a worker is the wage cost for the expected employment period of $1/b$, followed by $\bar{w}$ for the expected period of unemployment, $1/a$. This generates labor demand given by

$$f'(L_i) = w + \bar{w}b/(a+r).$$

[13] We have already shown that all firms offer the same employment benefits $\bar{w}$, so V_u is indeed a single number, i.e., an unemployed person's utility is independent of his previous employer.

Substituting the expression for V_u (i.e., (8)) into the *NSC* (5) yields the *aggregate NSC*

$$(9) \qquad w \geq \bar{w} + e + e(a+b+r)/q.$$

Notice that the critical wage for nonshirking is greater: (a) the smaller the detection probability q; (b) the larger the effort e; (c) the higher the quit rate b; (d) the higher the interest rate r; (e) the higher the unemployment benefit ($\bar{w}$); and (f) the higher the flows out of unemployment a.

We commented above on the first four properties; the last two are also unsurprising. If the unemployment benefit is high, the expected utility of an unemployed individual is high, and therefore the punishment associated with being unemployed is low. To induce individuals not to shirk, a higher wage must be paid. If a is the probability of obtaining a job per unit of time, $1/a$ is the expected duration of being unemployed. The longer the duration, the greater the punishment associated with being unemployed, and hence the smaller the wage that is required to induce nonshirking.

The rate a itself can be related to more fundamental parameters of the model, in a steady-state equilibrium. In steady state the flow *into* the unemployment pool is bL where L is aggregate employment. The flow *out* is $a(N-L)$ (per unit time) where N is the total labor supply. These must be equal, so $bL = a(N-L)$, or

$$(10) \qquad a = bL/(N-L).$$

Substituting for a into (9), the aggregate *NSC*, we have

$$(11) \quad w \geq e + \bar{w} + \frac{e}{q}\left(\frac{bN}{(N-L)} + r\right)$$
$$= e + \bar{w} + (e/q)(b/u + r) \equiv \hat{w},$$

where $u = (N-L)/N$, the unemployment rate. This constraint, the aggregate *NSC*, is graphed in Figure 1. It is immediately evident that *no shirking is inconsistent with full employment*. If $L = N$, $a = +\infty$, so any shirking worker would immediately be rehired. Knowing this, workers will choose to shirk.

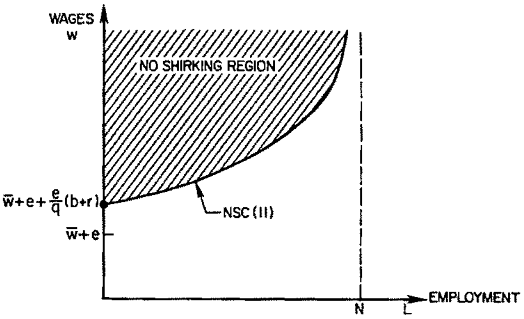


FIGURE 1. THE AGGREGATE NO-SHIRKING CONSTRAINT

The equilibrium wage and employment level are now easy to identify. Each (small) firm, taking the aggregate job acquisition rate a as given, finds that it must offer at least the wage $\hat{w}$. The firm's demand for labor then determines how many workers are hired at the wage. Equilibrium occurs where the aggregate demand for labor intersects the aggregate *NSC*. For $\bar{w} = 0$, equilibrium occurs when

$$F'(L) = e + (e/q)(bN/(N-L) + r).$$

The equilibrium is depicted in Figure 2.[14] It is important to understand the forces which cause E to be an equilibrium. From the firm's point of view, there is no point in raising wages since workers are providing effort and the firm can get all the labor it wants at w^*. Lowering wages, on the other hand, would induce shirking and be a losing idea.[15]

From the worker's point of view, *unemployment is involuntary*: those without jobs would be happy to work at w^* or lower, but cannot make a credible promise not to shirk at such wages.

[14] Aggregate labor demand is $F'(L)$ only when $\bar{w} = 0$ (see fn. 12).

[15] We have assumed that output is zero when an individual shirks, but we need only assume that a shirker's output is sufficiently low that hiring shirking workers is unprofitable.

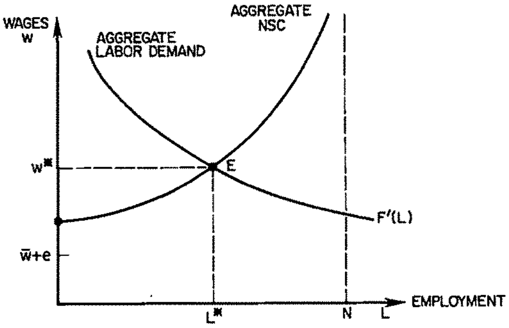


FIGURE 2. EQUILIBRIUM UNEMPLOYMENT

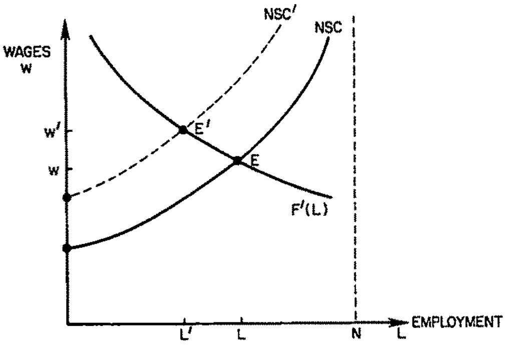


FIGURE 3. COMPARATIVE STATICS

Note: A decrease in the monitoring intensity q, or an increase in the quit rate b, leads to higher wages and more unemployment

Notice that the type of unemployment we have characterized here is very different from search unemployment. Here, all workers and all firms are identical. There is perfect information about job availability. There is a different information problem: firms are assumed (quite reasonably, in our view) not to be able to monitor the activities of their employees costlessly and perfectly.

E. *Simple Comparative Statics*

The effect of changing various parameters of the problem may easily be determined. As noted above, increasing the quit rate b, or decreasing the monitoring intensity q, decreases incentives to exert effort. Therefore, these changes require an increase in the wage necessary (at each level of employment) to induce individuals to work, that is, they shift the *NSC* curve upwards (see Figure 3). On the other hand, they leave the demand curve for labor unchanged, and hence the equilibrium level of unemployment and the equilibrium wage are both increased. Increases in unemployment benefits have the same impact on the *NSC* curve, but they also reduce labor demand as workers become more expensive, so they cause unemployment to rise for two reasons.

Inward shifts in the labor demand schedule create more unemployment. Due to the *NSC*, wages cannot fall enough to compensate for the decreased labor demand. The transition to the higher unemployment equilibrium will not be immediate: wage decreases by individual firms will only become attractive as the unemployment pool grows. This provides an explanation of wage sluggishness.

F. *Welfare Analysis*

In this section we study the welfare properties of the unemployment equilibrium. We demonstrate that the equilibrium is not in general Pareto optimal, when information costs are explicitly accounted for.

We begin with the case where the owners of the firms are the same individuals as the workers, and ownership is equally distributed among N workers. The central planning problem is to maximize the expected utility of the representative worker subject to the *NSC* and the resource constraint:

$$(12)\quad \max_{w,\bar{w},L}\,(w-e)L+\bar{w}(N-L)$$

subject to $w \geq e+\bar{w}+(e/q)((bN$

$$/(N-L))+r) \quad (NSC)$$

subject to $wL+\bar{w}(N-L)\leq F(L)$

(Feasibility)

subject to $\bar{w}\geq 0$.

Since workers are risk neutral it is easy to check[16] that the optimum involves $\bar{w}$ at the minimum allowable level, which is assumed to be 0. The reason is that increases in $\bar{w}$ tighten the *NSC*, so all payments should be made in the form of w rather $\bar{w}$.

Setting $\bar{w} = 0$, the problem simplifies to

$$(12') \quad \max_{w, L} (w - e)L$$

subject to $w \geq e + (e/q)((bN/(N-L)) + r)$;

and $\quad wL \leq F(L)$.

The set of points which satisfy the constraints is shaded in Figure 4. Iso-utility curves are rectangular hyperboles. So long as $F'(L) > e$, these are steeper than the average product locus, so the optimum occurs at point A where the *NSC* intersects the curve $w = F(L)/L$, that is, where wages equal the average product of labor. In contrast, the market equilibrium occurs at E where the marginal product of labor curve, $w = F'(L)$, intersects the *NSC* (Figure 2). Observe that in the case of constant returns to scale, $F'(L)L = F(L)$, so the equilibrium is optimal.

Wages should be subsidized, using whatever (pure) profits can be taxed away. An equivalent way to view the social optimum is a tax on unemployment to reduce shirking incentives; the wealth constraint on the unemployed requires that $\bar{w} \geq 0$, or equivalently that profits after taxes be nonnegative.[17] The optimum can be achieved by taxing away all profits and financing a wage subsidy of τ, shown in Figure 4. *The "natural" unemployment rate is too high.*

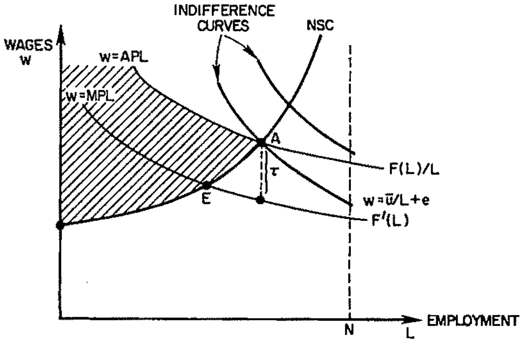


FIGURE 4. SOCIAL OPTIMUM AT *A*

In the case where the workers and the owners are distinct individuals, the tax policy described above would reduce profits, increase wages, and increase employment levels. While it would increase aggregate output (net of effort costs), such a tax policy would *not* constitute a Pareto improvement, since profits would fall. For this reason, the equilibrium is Pareto optimal in this case, even though it fails to maximize net national product. We thus have the unusual result that the Pareto optimality of the equilibrium depends upon the distribution of wealth. The standard separation between efficiency and income distribution does not carry over to this model.

It should not be surprising that the equilibrium level of unemployment is in general inefficient. Each firm tends to employ too few workers, since it sees the private cost of an additional worker as w, while the social cost is only e, which is lower. On the other hand, when a firm hires one more worker, it fails to take account of the effect this has on V_u (by reducing the size of the unemployment pool). This effect, a negative externality imposed by one firm on others as it raises its

[16] Formally,

$$\begin{aligned}\mathscr{L} &= (w-e)L + \bar{w}(N-L) \\ &\quad + \lambda[w - e - \bar{w} - (e/q)(bN/(N-l) + r)] \\ &\quad + \mu[F(L) - wL - \bar{w}(N-L)].\end{aligned}$$

Differentiating with respect of w and $\bar{w}$ yields

$$\mathscr{L}_w = L + \lambda - \mu L \leq 0 \text{ and} = 0 \text{ if } w > 0.$$

$$\mathscr{L}_{\bar{w}} = (N-L) - \lambda - \mu(N-L) \leq 0 \text{ and} = 0 \text{ if } \bar{w} > 0.$$

We know $w > 0$ by the *NSC*, so $\mathscr{L}_w = 0$, i.e., $L(1-\mu) + \lambda = 0$. Therefore, since $\lambda > 0$, $\mu > 1$. But then $\mathscr{L}_{\bar{w}} = (N - L)(1-\mu) - \lambda < 0$. This implies that $\bar{w} = 0$.

[17] The constraint $\bar{w} \geq 0$ can be rewritten, using the resource constraint, as $F(L) - wL \geq 0$, i.e., $\pi \geq 0$.

level of employment, tends to lead to overemployment. In the simple model presented so far, the former effect dominates, and the natural level of unemployment is too high. This will not be true in more general models, however, as we shall see below.

II. Extensions

In this section we describe how the results derived above are modified or extended when we relax some of the simplifying assumptions. We discuss three extensions in turn: endogenous monitoring, risk aversion, and endogenous turnover. Detailed derivations of the claims made below are available in our earlier working paper.

A. *Endogenous Monitoring*

When employees can select the monitoring intensity q, they can trade off stricter monitoring (at a cost) with higher wages as methods of worker discipline. In general, firms' monitoring intensities will not be optimal, due to the externalities between firms described above. In general, it is not possible to ascertain whether the equilibrium entails too much or too little employment. In the case of constant returns to scale ($F(L)=L$), however (which led to efficiency with exogenous monitoring), the competitive equilibrium involves too much monitoring and too much employment.

The result is not as unintuitive as it first seems: each firm believes that the only instrument at its control for reducing shirking is to increase monitoring. There is, however, a second instrument: by reducing employment, workers are induced not to shirk. This enables society to save resources on monitoring (supervision). These gains more than offset the loss from the reduced employment.

It is straightforward to see how this policy may be implemented. If firms can be induced to reduce their monitoring, welfare will be increased. Hence a tax on monitoring, with the proceeds distributed, say, as a lump sum transfer to firms, will leave the no-shirking constraint/national-resource constraint unaffected, but will reduce monitoring.

B. *Risk Aversion*

With risk neutrality, the optimum and the market both involve $\bar{w}=0$. Clearly $\bar{w}=0$ cannot be optimal if workers are highly risk averse and may be separated from their jobs for exogenous reasons. Yet the market always provides $\bar{w}=0$ (or the legal minimum). The proof above that $\bar{w}=0$ carries over to the case of risk-averse workers.

When equilibrium involves unemployment, firms have no difficulty attracting workers and hence offer $\bar{w}=0$, since $\bar{w}>0$ merely reduces the penalty of being fired. When *other* firms offer $\bar{w}=0$, this argument is only strengthened: unemployed workers are even easier to attract. It is striking that the market provides no unemployment benefits even when workers are highly risk averse. Clearly the social optimum involves $\bar{w}>0$ if risk aversion is great enough. This may provide a justification for mandatory minimum benefit levels.

C. *Endogenous Turnover*

In general a firm's employment package will influence the turnover rate it experiences among its employees. Since the turnover rate b affects the rate of hiring out of the unemployment pool, and hence V_u, it affects other firms' no-shirking constraints. Because of this externality, firms' choices of employment packages will not in general be optimal. This type of externality is similar to search externalities in which, for example, one searcher's expected utility depends on the number or mix of searchers remaining in the market. In the current model, policies which discourage labor turnover are attractive as they make unemployment more costly to shirkers.

III. Alternative Methods for the Enforcement of Discipline

This paper has explored a particular mechanism for the enforcement of discipline: individuals who are detected shirking are fired, and in equilibrium the level of unemployment is sufficiently large that this threat serves as an effective deterrent to shirking. The question naturally arises whether there are alternative, less costly, or more effective discipline mechanisms.

A. *Performance Bonds*

The most direct mechanism by which discipline might be enforced is through the posting by workers of performance bonds. Under this arrangement the worker would forfeit the bond if the firm detected him shirking. One problem with this solution is that workers may not have the wealth to post bond.[18] A more fundamental problerh with this mechanism is that the firm would have an incentive to *claim* that the worker shirked so that it could appropriate the bond. Assuming, quite realistically, that third parties cannot easily observe workers' effort (indeed, it is usually more costly for outsiders to observe worker inputs than for the employer to do so), there is no simple way to discipline the *firm* from this type of opportunism.

Having recognized this basic point, it is easy to see that a number of other plausible solutions face the same difficulty. For example, consider an employment package which rewards effort by raising wages over time for workers who have not been found shirking. This is in fact equivalent to giving the worker a level wage stream, but taking back part of his earlier payments as a bond, which is returned to him later. Therefore, by the above argument, the firm will have an incentive to fire the worker when he is about to enter the "payoff" period in which he recovers his bond. This is the equivalent to the firm's simply appropriating the bond. It is optimal for the firm to replace expensive senior workers by inexpensive junior ones.[19]

Clearly the firm's reputation as an honest employer can partially solve this problem; the employer is implicitly penalized for firing a worker if this renders him less attractive to prospective employees. Yet this reputation mechanism may not work especially well, since prospective employees often do not know the employer's record, and previous dismissals may have been legitimate (it is not possible for prospective employees to distinguish legitimate from unfair earlier dismissals, if they are aware of them at all). If the reputation mechanism is less than perfect, it will be augmented by the unemployment mechanism.

B. *Other Costs of Dismissal*

Unemployment in the model above serves the role of imposing costs on dismissed workers. If other costs of dismissal are sufficiently high, workers may have an incentive to exert effort even under conditions of full employment. Examples of such costs are search costs, moving expenses, loss of job-specific human capital, etc. In markets where these costs are substantial, the role of equilibrium unemployment is substantially diminished. The effect we have identified above will still be present, however, when effort levels are continuous variables: each firm will still find that employee effort is increasing with wages, so wages will be bid up somewhat above their full-employment level. The theory predicts that involuntary (as well as frictional) unemployment rates will be higher for classes of workers who have lower job switching costs.

[18] This is especially true if detection is difficult (low q) so that an effective bond must be quite large. Even if workers could borrow to post the bond, so long as bankruptcy is possible, the incentives for avoiding defaulting on the bond are not different from the incentives to avoid being caught shirking by the firm in the absence of a bond. Note once again the importance of the wealth distribution in determining the nature of the equilibrium. If all individuals inherit a large amount of wealth, then they could post bonds.

[19] In competitive equilibrium, the average (discounted) value of the wage must be equal to the average (discounted) value of the marginal product of the worker. If there is a bonus for not shirking, *initially* the wage must be below the value of the marginal product. It is as if the worker were posting a bond (the difference between his marginal product and the wage), and as such this scheme is susceptible to precisely the same objections raised against posting performance bondings. The employer has an incentive to appropriate the bond. Since workers know this, this is not a viable incentive scheme. For a fine study in which firms' reputations are assumed to function so as to make this scheme viable, see Edward Lazear (1981).

C. *Heterogeneous Workers*

The strongest assumption we have made is that of identical workers. This assumption ruled out the possibility that firing a worker would carry any stigma. Such a stigma could serve as a discipline device, even with full employment.[20] In reality, of course, employers *do* make wage offers which are contingent on employment history. Such policies make sense when firms face problems of adverse selection.

We recognize that workers' concern about protecting their reputations as effective, diligent workers may provide an effective incentive for a disciplined labor force.[21] Shapiro's earlier (1983) analysis of reputation in product markets showed, however, that for reputations to be an effective incentive device, there must be a cost to the loss of reputation. It is our conjecture that, under plausible conditions, even when reputations are important, equilibrium will entail some use of unemployment as a discipline device for the labor force, at least for lower-quality workers. An important line of research is the study of labor markets in which adverse selection as well as moral hazard problems are present. In this context, our model should provide a useful complement to the more common studies of adverse selection in labor markets.

IV. Conclusions

This paper has explored the role of unemployment, or job rationing, as an incentive device. We have argued that when it is costly to monitor individuals, competitive equilibrium will be characterized by unemployment, but that the natural rate of unemployment so engendered will not in general be optimal. We have identified several forces at work, some which tend to make the market equilibrium unemployment rate too high, and others which tend to make it too small. Each firm fails to take into account the consequences of its actions on the level of monitoring and wages which other firms must undertake in order to avoid shirking by workers. Although these externalities are much like pecuniary externalities, they are important, even in economies with a large number of firms.[22] As a result, we have argued that there is scope for government interventions, both with respect to unemployment benefits and taxes or subsidies on monitoring and labor turnover, which can (if appropriately designed) lead to Pareto improvements.

The type of unemployment studied here is not the only or even the most important source of unemployment in practice. We believe it is, however, a significant factor in the observed level of unemployment, especially in lower-paid, lower-skilled, blue-collar occupations. It may well be more important than frictional or search unemployment in many labor markets.

[20] See Bruce Greenwald (1979) for a simple model in which those who are in the "used labor market" are in fact a lower quality than those in the "new" labor market.

[21] This suggests once again that our results may be most significant in labor markets for lower-quality workers: in such markets employment histories are utilized less and workers already labeled as below average in quality have less to lose from being labeled as such.

[22] For a more general discussion of pecuniary, or more general market mediated externalities, with applications to economies with important adverse selection and moral hazard problems, see Greenwald and Stiglitz (1982).

REFERENCES

Alchian, Armen A. and Demsetz, Harold, "Production, Information Costs, and Economic Organization," *American Economic Review*, December 1972, *62*, 777–95.

Calvo, Guillermo A., "Quasi-Walrasian Theories of Unemployment," *American Economic Review Proceedings*, May 1979, *69*, 102–06.

______, "On the Inefficiency of Unemployment," Columbia University, October 1981.

______ **and Wellisz, Stanislaw,** "Hierarchy, Ability and Income Distribution," *Journal of Political Economy*, October 1979, *87*, 991–1010.

Diamond, Peter, "Mobility Costs, Frictional Unemployment, and Efficiency," *Journal* *of Political Economy*, August 1981, *89*, 798–812.

Greenwald, Bruce, C. N., *Adverse Selection in the Labor Market*, New York; London: Garland, 1979.

______ **and Stiglitz, Joseph E.,** "Pecuniary Externalities," unpublished, Princeton University, 1982.

Lazear, Edward P., "Agency, Earnings Profiles, Productivity, and Hours Restrictions," *American Economic Review*, September 1981, *71*, 606–20.

Salop, Steven C., "A Model of the Natural Rate of Unemployment," *American Economic Review*, March 1979, *69*, 117–25.

Shapiro, Carl, "Premiums for High Quality Products as Returns to Reputations," *Quarterly Journal of Economics*, November 1983, *98*, 658–79.

______ **and Stiglitz, Joseph E.,** "Equilibrium Unemployment as a Worker Discipline Device," Discussion Papers in Economics, No. 28, Woodrow Wilson School, Princeton University, April 1982.

Stiglitz, Joseph E., "Prices and Queues as Screening Devices in Competitive Markets," IMSSS Technical Report No. 212, Stanford University, 1976.

______ **and Weiss, Andrew,** "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, *71*, 393–410.

Weiss, Andrew, "Job Queues and Layoffs in Labor Markets with Flexible Wages," *Journal of Political Economy*, June 1980, *88*, 526–38.

Wilson, Charles, "The Nature of Equilibrium in Markets with Adverse Selection," *Bell Journal of Economics*, Spring 1980, *11*, 108–30.

Farm Foreclosure Moratorium Legislation: A Lesson from the Past

By LEE J. ALSTON*

The recent protests over farm foreclosures and the calls for legislation to prevent farm foreclosures are not without precedent. Farm foreclosures were high throughout the 1920's and 1930's, in large part because farmers had to make mortgage payments fixed in nominal dollars while their earnings were falling.[1] In response, many state legislatures sought to aid farmers by passing legislation that postponed foreclosure by creditors. Scholars have given little attention to these various state emergency measures and their impact on creditors and debtors.[2] This article is a step toward filling this void. The first section contains a general description of the moratorium legislation, and hypotheses about the causes of moratorium legislation are proposed and tested. In Section II, the economic consequences of moratorium legislation are addressed and the hypotheses are formally specified and tested.

I. Moratorium Legislation: Causes

Legislation to relieve the burden of mortgage debt in hard times originated before the 1930's. As early as 1820, the legislature of New York passed a statute giving debtors a year of grace before land could be sold (Lawrence Friedman, 1973, p. 217). Other states passed similar laws extending the redemption period at various times during the nineteenth century, generally in the wake of a recession.[3] What differentiates the emergency debt legislation in the nineteenth century from the laws in the 1930's was the pervasiveness of state legislation in the 1930's and the role played by the courts.

Legislatures in both the nineteenth century and in the 1930's appear to have been responsive to debtors' demands during economic downturns. In the nineteenth century, however, the courts generally declared debtor relief legislation unconstitutional as a violation of the obligation of contracts, but, in 1934, in *Home Building and Loan Association v. Blaisdell et al.* (hereafter, *Home BLA v. Blaisdell*), the U.S. Supreme Court, in a five-to-four decision, upheld a farm foreclosure moratorium enacted in Minnesota (Friedman, pp. 215–18, and Robert Skilton). The dissenting justices argued, as did the judges in the nineteenth century,

> That the contract impairment clause denies to the several states the power to mitigate hard consequences resulting to debtors from financial or economic exigencies by an impairment of the obligation of contracts of indebtedness. A candid consideration of the history and circumstances which led up to and accompanied the framing and adoption of this clause will demonstrate conclusively that it was framed and adopted with the specific and studied purpose of preventing legislation designed to relieve debtors *especially* in time of financial distress. Indeed, it is not probable that any other purpose was definitely in the minds of those who

*Department of Economics, Williams College, Williamstown, MA 01267. For comments on earlier drafts, I thank Mary Alston, Roger Bolton, Ralph Bradburd, Barry Eichengreen, Price Fishback, Michael Fortunato, Robert Higgs, Michael Horrigan, Michael McPherson, Thomas Morawitz, Richard Parks, Randy Rucker, Andrew Rutten, Morton Schapiro, John Sheahan, Peter Temin, Richard Valelly, Carl Van Duyne, an anonymous referee, and participants at seminars at Harvard University, Williams College, and the University of Washington. The initial research on this paper was undertaken while I was on leave at the University of Washington. I am grateful for their generous support.

[1] For a discussion of the causes of farm distress, see my 1983 article.

[2] A notable exception is Archibald Woodruff (1937).

[3] See Friedman (pp. 215–18; 324–25), and Robert Skilton (1943). Skilton cites and discusses the various important state statutes and judicial decisions concerning redemption periods.

> composed the framers' convention or the ratifying state conventions which followed, although the restriction has been given a wider application upon principles clearly stated by Chief Justice Marshall in the Dartmouth College Case.
>
> [*Home BLA v. Blaisdell*, pp. 453–54]

The majority of the justices argued in upholding the judgment of the Supreme Court of Minnesota that "while emergency does not create power, emergency may furnish the occasion for the exercise of power" (*Home BLA v. Blaisdell*, p. 426). In essence the court ruled that "the economic interests of the State may justify the exercise of its continuing and dominant protective power notwithstanding interference with contracts" (*Home BLA v. Blaisdell*, p. 437). Although the judgment of the court in 1934 differed from the decisions in the nineteenth century, the decision could have had little impact on the impetus for legislation, because by the time of the courts' decision twenty-two of the twenty-five farm foreclosure moratoria had been enacted.[4]

Although mortgage debt relief legislation had historical roots, never was it so widespread across states as in the 1930's. From 1932 to 1934, farm foreclosure moratorium legislation was enacted in twenty-five states. (See Table 1.) These laws prevented creditors, in the event of default by a debtor, from obtaining title to the land for a specified or court-determined period of time. The allowable moratorium period varied by state from 3 months to almost 4 years.[5] During the moratorium period, debtors were usually obligated to pay a rental price that was generally set by the state courts below the free market rental (Archibald Woodruff, p. 114). In addition, creditors received only part of the rental payments, as taxes were deducted before creditors were paid. Creditors retained limited rights to foreclose if nonpayment of the rental price or wastage of property occurred.

TABLE 1—STATE MORATORIUM LEGISLATION

States Passing Legislation	States Not Passing Legislation
Arizona	Alabama
California	Arkansas
Delaware	Colorado
Idaho	Connecticut
Illinois	Florida
Iowa	Georgia
Kansas	Indiana
Louisiana	Kentucky
Michigan	Maine
Minnesota	Maryland
Mississippi	Massachusetts
Montana	Missouri
Nebraska	New Jersey
New Hampshire	New Mexico
New York	Nevada
North Carolina	Oregon
North Dakota	Rhode Island
Ohio	Tennessee
Oklahoma	Utah
Pennsylvania	Virginia
South Carolina	Washington
South Dakota	West Virginia
Texas	Wyoming
Vermont	
Wisconsin	

Sources: U.S. Congress, Central Housing Committee; USDA, Bureau of Agricultural Economics, "State Measures for the Relief of Agricultural Indebtedness in the United States, 1932 and 1933"; and "State Measures..., 1933 and 1934."

Federal legislation dealing with agricultural distress, rather than state legislation, has been the primary concern of most scholars researching the New Deal period.[6] Per-

[4] The potentially damaging long-run implications of the decision reached in *Home BLA v. Blaisdell* were forcefully argued as follows by Justice Sutherland in his dissent.

> He simply closes his eyes to the necessary implications of the decision who fails to see in it the potentiality of future gradual but ever-advancing encroachments upon the sanctity of private and public contracts. The effect of the Minnesota legislation, though serious enough in itself, is of trivial significance compared with the far more serious and dangerous inroads upon the limitations of the Constitution which are almost certain to ensure as a consequence naturally following any step beyond the boundaries fixed by that instrument.
>
> [p. 448].

Future research will analyze the extent to which Justice Sutherland's prediction was valid by an examination of cases involving an abridgement of contracts. Particular attention will be given to the precedents cited.

[5] See sources cited in Table 1.

[6] It is interesting to compare the different reactions by the federal and state governments to agricultural dis-

haps their emphasis is justified in terms of the long-run consequences of federal intervention in agriculture, but nevertheless their cursory treatment of the state legislation frequently gives a misleading impression of its causes. Historians have tended to focus on the violence that preceded moratorium legislation in a few states, and by omission have given little sense of the ubiquity of such legislation, or of what, besides violence, prompted twenty-five states to enact moratorium laws. Typically, however, those authors, who have most thoroughly researched either farm protest or moratorium legislation, tend to discount the causal role played by violence. The following quotations, from sources beginning with general historical accounts of the depression and ending with a special report to a congressional committee, illustrate this generalization.

> The attempt to boost farm prices failed, but the farmers used still more violent means to halt foreclosures. At Storm Lake, Iowa, rope-swinging farmers came close to hanging a lawyer conducting a foreclosure; in the Le Mars area, five hundred farmers marched on the courthouse steps and mauled the sheriff and the agent of a New York mortgage company. "If we don't get beneficial service from the Legislature," warned a leader of the Nebraska holiday movements, "200,000 of us are coming to Lincoln, and we'll tear that new State Capital Building to pieces."
> [William Leuchtenburg, 1963, p. 24]

> In other instances, crowds of farmers intimidated judges, sheriffs and their deputies, or agents of the creditors, thereby preventing the sales entirely.... The legislature of a number of states responded by passing moratorium laws.... [Van Perkins, 1969, p. 16]

> The legislation passed by states and Federal Government in early 1933 was undoubtedly much influenced by the unrest. [Woodruff, 1937, p. 106].

> Moratorium laws would probably have been enacted eventually without grassroots agitation, but penny auctions and marches to the capital hastened and placed the "urgent" label on the political responses.... The passage of moratorium laws reflected upon the Farmers' Holiday Association more than any other single farm organization, giving it an illusion of strength greater than it possessed.
> [John Shover, 1965, p. 88]

> There is no question but that such emergency laws have arisen because depressed economic conditions prevailing throughout the country greatly increased the number of mortgagors who were unable to fulfill their contracts. [U.S. Congress, Central Housing Committee, 1936, p. 2]

There is no doubt that in some states debt-burdened farmers brought considerable pressure to bear on their legislators, but not all pressure was in the form of violence. Violence was the extreme manifestation of pressure and erupted, to my knowledge, in only a few states, most of them in the Midwest.[7] Economic distress, being more widespread, was more likely the underlying cause of legislation and protest.

To conceptualize how economic distress evoked moratorium legislation, we need a model of the legislative process. The simple model tested below assumes that politicians, in order to be reelected, enact legislation that satisfies the demands of the electorate.[8] Demand in this model determines the quantity of legislation. In this particular instance, the demand for moratorium legislation is hy-

tress. The federal government tended to intervene directly in markets through quotas, price supports, and refinancing of mortgage loans through the federal land banks. States did not have this option, because agricultural markets extended beyond state boundaries. Hence, it seems natural that moratoria occurred at the state level. Whether these differing reactions would apply to other crises would be an informative research inquiry.

[7] States in which violence erupted include Iowa, Minnesota, North Dakota, Pennsylvania, South Dakota, and Wisconsin (Skilton).

[8] Creditors are also conceivably demanders or resisters of legislation but, given their spatial and institutional diversity in the 1930's, collusion appears to have been prohibitively costly (see my forthcoming article).

TABLE 2—REGRESSION RESULTS

Equations	Constant	Foreclosures per 1,000 Farms	Farms Mortgaged	Federally Held Mortgage Debt
1	−1.40	0.08	0.01	−0.04
		(2.31)[a]	(0.26)	(−1.59)[b]
2	−1.23	0.11	0.01	−0.06
		(2.27)[a]	(0.24)	(−2.00)
3	−1.66	0.10	0.03	−0.05
		(1.99)[a]	(0.57)	(−1.66)[b]
4	−0.95	0.10	0.02	−0.06
		(2.20)[a]	(0.63)	(−2.04)[a]

Sources: Moratorium legislation: see Table 1. Foreclosures..., USDA, "The Farm Real Estate Situation, 1933–34"; Farms Mortgaged = percent of total owner-operated farms reported as mortgaged in 1930, *Fifteenth Census of the United States, 1930: Agriculture,* Vol. II, Table 18; Federally Held Mortgaged Debt = the sum of federal land bank loans from 1917 to 1932 divided by the total farm mortgage debt in 1932, farm mortgage debt is found in USDA, "Farm Mortgage Credit Facilities in the United States," Table 64, p. 221, as are data on federal land bank loans, Table 78, p. 245.

Notes: Dependent variable = 1 if moratorium legislation enacted, 0 otherwise. The *t*-ratios are shown in parentheses.

[a]Significant at the 95 percent confidence level.

[b]Significant at the 90 percent confidence level.

pothesized to be a positive function of economic distress, with economic distress measured by a proxy variable: the average number of farm foreclosures per 1,000 farms in 1932 and 1933 in each state. In addition to considering the actual number of foreclosures, farmers may have based their demand for a moratorium on their expectations of future foreclosures. Two proxy variables capture this aspect of demand: the percentage of farms mortgaged in a state and the percentage of mortgage debt held by federal land banks. The more farms were mortgaged, the more farms potentially could be foreclosed, and therefore the greater the number of demanders of moratorium legislation. On the other hand, the greater the percentage of mortgage debt held by federal land banks, the fewer were the demanders for moratorium legislation, because federal land banks were less likely than other creditors to foreclose (Woodruff, p. 104). Farmers in debt to the federal land banks would feel less threatened and therefore would exert less pressure on their legislators to enact a moratorium.

To test the hypotheses presented above, a logistic model was employed to estimate the parameters of a regression equation in which moratorium legislation in each state is predicted to be positively influenced by the number of farm foreclosures and percentage of farms mortgaged and negatively influenced by the amount of mortgage debt held by the federal land banks.[9] The results, presented in Table 2, are consistent with two of the three hypotheses. The data set, in equation 1, contains observations on all forty-eight states. The dependent variable was constructed by determining whether or not a state passed moratorium legislation. The coefficients on both the foreclosure and federally held mortgage debt variables have the predicted signs and are significant at the 95 and 90 percent confidence levels, respectively. The coefficient on the percentage of farms mortgaged has the predicted sign, but is not statistically significant.

The model estimated in equation 1 may be misspecified. Whether or not a state enacted moratorium legislation may have depended on the leniency of existing laws concerning foreclosure. The author of the special con-

[9] The parameters (βi) of the linear logistic model estimated are $E(Y) = \theta = e^{\beta X}/(1 + e^{\beta X})$, where Y denotes the dependent variable and X denotes the independent variables.

gressional report on moratorium legislation commented on the effect in certain states of existing laws:

> As foreclosure laws are lenient on mortgagor, providing for a long period in which mortgagor may pay before decree given and as mortgagor may redeem at any time within ten years from last recognition by mortgagee of right of redemption, *the Legislature of this State* [Georgia] *evidently determined that no emergency mortgage moratorium legislation was necessary.*
> [Central Housing Committee, p. A-6; emphasis added]

The author comments on the leniency of foreclosure laws in seven other states, in addition to Georgia, though he is not usually so definite in his opinion of the preventative effect existing laws had on the passage of moratoria. For example, in the case of Washington, he notes: "No emergency mortgage moratorium legislation has been enacted in this state. However redemption provisions of present foreclosure laws are rather lenient" (Central Housing Committee, p. A-26). The eight states explicitly cited as having no moratorium but lenient foreclosure laws are Connecticut, Georgia, Indiana, Massachusetts, Missouri, Nevada, Tennessee, and Washington.

To determine whether the results in equation 1 are sensitive to my classification, I estimated equation 2, omitting the above eight states from my sample. The results are similar to those of equation 1. Again, the coefficients on the foreclosure and federally held mortgage debt variables have the predicted signs and are significant at the 95 percent confidence level. The coefficient on the percentage of farms mortgaged remains insignificant.

In addition to commenting on the leniency of legislation in certain states, the author of the special congressional report notes that in three states—Utah, Virginia, and Wyoming—execution of the foreclosure was within the discretion of the court. I have no evidence that this discretion was regularly employed, but on the assumption that existing laws in these three states discouraged enactment of moratoria, I estimated equation 3 with a data set that omitted Utah, Virginia, and Wyoming, plus the states excluded in equation 2. The results differ little from those of equations 1 and 2. The coefficients of the foreclosure and federally held debt variables remain significant at the 95 and 90 percent confidence levels, respectively.

One might speculate that moratorium legislation would have been passed in the absence of existing legislation that was lenient to debtors in the above eight states mentioned in the congressional report. On the basis of this counterfactual conjecture, I estimated equation 4 on a sample that includes all forty-eight states and assumes that Connecticut, Massachusetts, Georgia, Indiana, Missouri, Nevada, Tennessee, and Washington enacted moratorium legislation. The results are similar to those of the previous equations.[10] It appears that the model presented to explain the enactment of foreclosure relief is quite robust with respect to my classification of marginal cases.

The model specified in Table 2 assumes that the likelihood of moratorium legislation depends on the severity of farm distress, but is not influenced by the importance of agriculture in each state. If political pressure is partly determined by the resources available to effectively lobby Congress, then equations 1–4 in Table 2 have an omitted variable. To test an alternative model, I added to the regression equations the ratio of agricultural income to total state income as a measure of the potential political-economic strength of farmers.[11] Again, a logistic model was used to estimate the parameters of the regression equations using the same data set as reported in Table 2. In general, the results presented

[10] An additional equation was estimated in which Virginia, Utah, and Wyoming, in addition to the eight "quasi-moratorium" states in equation 4, were assumed to have passed moratorium legislation. The results are similar to the others reported in Table 2. The coefficients of the foreclosure and federally held debt variables are significant at the 95 and 90 percent confidence levels, respectively, and the coefficient on percentage of farms mortgaged remains insignificant.

[11] I am grateful to the referee for suggesting this alternative specification.

TABLE 3—REGRESSION RESULTS

Equations	Constant	Foreclosures Per 1,000 Farms	Farms Mortgaged	Federally Held Mortgage Debt	Agricultural Income/ Total State Income
1	−.25	.04	.002	−.07	.19
		(1.06)	(.05)	(−1.99)[a]	(1.51)[b]
2	−.64	.09	.004	−.07	.09
		(1.69)[b]	(.11)	(−1.99)[a]	(.66)
3	−1.16	.08	.02	−.06	.09
		(1.43)[b]	(.50)	(−1.69)[a]	(.61)
4	−.76	.10	.02	−.07	.05
		(1.78)[a]	(.04)	(−2.01)[a]	(.36)

Sources: Foreclosures..., Farms Mortgaged, and Federally Held Mortgage Debt: see Table 2. Agricultural Income is constructed multiplying state per capita farm income in 1930 by the farm population in 1930. Data on per capita farm and personal income by states are found in Frank Hanna (1959, Table 61, p. 248, and Table 1, p. 28). Data on farm population and total population in 1930 by states are found in *Statistical Abstract of the United States—1932*, Table 36, p. 47, and Table 11, pp. 8–9.

Notes: Dependent variable = 1 if moratorium legislation enacted. The *t*-statistics are shown in parentheses.

[a]Significant at the 95 percent confidence level.

[b]Significant at the 90 percent confidence level.

in Table 3 do not support the hypothesis concerning the importance of the agricultural sector. Only in equation 1 is the empirical result weakly consistent with the hypothesis. Including the ratio of agricultural income to total state income in the regression equations reduces somewhat the statistical significance of the other independent variables but the coefficients change only slightly from those in Table 2.[12] My rationale for presenting both sets of regression results is to allow comparisons of the two models of the legislature process.

Since a moratorium was an abridgement of the enforceability of contracts, one might speculate that states might differ on ideological grounds *ceteris paribus* as to the merits of government intervention in private markets. A conservative political ideology could be thought of as influencing either the demanders of legislation—farmers would be less apt to petition for relief—or the suppliers of legislation—legislators would be less likely to give relief. To the extent that southern states were more averse than other states to government intervention, adding a dummy variable to the regressions should capture ideological differences as well as other regional differences. This exercise was performed with no support for the hypothesis that southern states differed from other states in the likelihood of passage of moratorium legislation.[13]

The simple model of legislation depending on political pressure, which in turn was dependent on the severity of economic distress and expectations of future distress, is highly consistent with the test results in Tables 2 and 3. The hypotheses had several opportunities for refutation, being tested against four classifications of the legislative output.

[12] Political pressure may partly be determined by the number of people who would most likely benefit from a moratorium on farm foreclosures. For this reason a variable capturing this aspect of the importance of agriculture was substituted in equations 1–4 for the ratio of agriculture income to state income. The variable, the percentage of workers engaged in agriculture, was constructed by dividing the number of males and females engaged in agriculture in each state by the number of gainful workers in each state. The empirical results do not support this measure of the importance of agriculture. In none of the equations is the coefficient significantly different from zero. The inclusion of this variable has little effect on the magnitude or significance of the other coefficients in the regression equation.

[13] I am grateful to the referee for suggesting that I test for ideological differences between the South and the rest of the nation.

II. The Economic Consequences of Moratorium Legislation

State legislatures passed moratoria on farm foreclosures in an attempt to prevent farmers from losing their farms. Were the moratoria successful? Those scholars who have commented on this issue argue that moratoria postponed some foreclosures but most likely averted few (Perkins, p. 16, and Woodruff, p. 104). In light of the low farm earnings at the time of legislation, most creditors were already quite lenient on delinquent debtors, provided farmers made reasonable efforts to pay and did not abuse the property. Creditors tended not to foreclose on owner-operated farms (Theodore Saloutos and John Hicks, 1961, p. 448, and Woodruff, p. 104). The implication is that any "temporary" halt in foreclosures would not provide enough time for earnings to increase to levels sufficient to avoid foreclosure on the types of loans actually being foreclosed in 1932. Most of the evidence on the leniency of creditors pertains to insurance companies and federal financial institutions. These groups however held only 12.8 percent of the farm mortgages in 1932, while individuals held 45 percent of all farm mortgages nationwide.[14] We would expect that individuals—and perhaps local banks—being less diversified and facing a more severe income constraint than insurance companies and federal financial institutions had more of an incentive to foreclose. In addition, insurance companies may have felt a greater need to maintain goodwill because of the historical distrust of eastern lending institutions. If different foreclosure practices existed across creditors, this would better explain the pressure by debtors for moratorium legislation.[15] Evidence on this issue would help determine the short- and long-run benefits to debtors. The impact on creditors however was certainly negative. The contractual rights of lenders were impaired by the temporary loss of the option to foreclose. Losses resulted from deterioration of farms from neglect or malice and from the courts' setting of rents below what creditors could obtain in a market (Woodruff, p. 114). We would expect that creditors, in order to protect themselves from future losses, would take measures that would either reduce the probability of delinquency on new loans or compensate them for losses from delinquencies.

Creditors could take two measures to reduce the harmful effects of a moratorium on farm foreclosures. First, creditors could increase interest rates on all new loans in order to compensate themselves for the losses incurred because of their inability to foreclose delinquent loans. This form of protection is likely to occur only if all loans are equally likely to become delinquent. If lenders had some *ex ante* information on what type of farmer was more likely to default on payments, they would increase interest rates on loans to this more delinquency-prone group of farmers. In this event we should observe relatively higher interest rates and fewer loans in states passing moratoria because some farmers would be discouraged from taking a loan at the increased interest rates. Second, or alternatively, to reduce the number of defaults, eligibility requirements could be tightened. Loans would be made to farmers possessing attributes associated with fewer defaults in the past.

Creditors may have been less likely to raise interest rates than to ration credit, based on the probability of default of different borrowers. Raising interest rates in times of unrest on the farm may generate increased hostility and ill will. In addition, since all of the moratoria were temporary, the foregone profits of making a wrong estimate of the ill will generated by raising interest rates appears minimal.

We are interested in determining the effects of moratorium legislation on the quantity of loans supplied and on the interest rates in the private mortgage market in each state. The number of private mortgage loans and equilibrium interest rates in each state are determined by both the demand for private

[14] See U.S. Department of Agriculture (1942, p. 224). For a discussion of the spatial and temporal variance in the importance of creditors see my forthcoming article.

[15] Unfortunately I have not been able to find any empirical evidence on this issue with the exception that generally federal institutions were less likely to foreclose.

and government mortgage loans and the supply of loans. The demand and supply equations for each state can be expressed as

$$(1)\quad Q_G^S = \alpha_0 + \alpha_1 E,$$

$$(2)\quad Q_P^S = \beta_0 + \beta_1 M + \beta_2 i + \beta_3 S + \beta_4 E,$$

$$(3)\quad Q_G^D = \overline{Q}_G^S \quad \text{if } i_G < i_p,$$

$$(4)\quad Q_P^D = \gamma_0 + \gamma_1 i + \gamma_2(\alpha_0 + \alpha_1 E),$$

where Q_G^S = supply of mortgage loans by the government, Q_P^S = supply of mortgage loans by private creditors,[16] Q_G^D = demand for government mortgage loans, Q_P^D = demand for mortgage loans from private creditors, E = expected farm earnings, M = moratorium, i = interest rate, and S = size of loan.

In the government's supply equation, $\alpha_1 < 0$. After 1932, the federal government made loans across states primarily on the basis of need. Relatively high farm earnings in a state induced the government to supply fewer loans.

In the private sector's supply equation: a moratorium, by eliminating the option to foreclose, imposes losses on creditors, who respond by supplying fewer loans; hence $\beta_1 < 0$. Since interest is the compensation creditors receive for making loans, the higher the interest rate, the more loans will be supplied; hence $\beta_2 > 0$. The transactions costs (negotiation, supervision, and enforcement) of a mortgage loan do not vary substantially across loan size and to this extent the costs per dollar of supplying credit fall the larger the loan. Creditors therefore are willing to supply more loans at every given interest rate the larger the average loan size; hence $\beta_3 > 0$. Greater expected farm earnings reduce creditors' expectations of defaults and thereby reduce their expectations of costs. Creditors respond by increasing their supply of loans at every interest rate; hence $\beta_4 > 0$.

In the demand for government loans equations, private borrowers are assumed to accept as many loans as the government issues, provided the interest rate on government loans is less than the prevailing rate on private mortgage loans. This was the case with the advent of the New Deal. In the demand for private loans equation: since interest is the cost of a loan, the higher the interest rate the fewer loans demanded; hence $\gamma_1 < 0$. The more loans the government supplies, the less the demand for loans from the private sector; hence $\gamma_2 < 0$.

The equilibrium quantity of private loans (Q_P) and equilibrium interest rate (i) are determined by the interaction of the supply and demand for private mortgage loans. The reduced-form solutions for Q_P and i are

$$(5)\quad Q_P = [(\beta_2\gamma_0 + \beta_2\gamma_2\alpha_0 - \gamma_1\beta_0) - \beta_1\gamma_1 M - \beta_3\gamma_1 S + (\alpha_1\beta_2\gamma_2 - \beta_4\gamma_1)E]/(\beta_2 - \gamma_1);$$

$$(6)\quad i = [(\gamma_0 + \gamma_2\alpha_0 - \beta_0) - \beta_1 M - \beta_3 S + (\gamma_2\alpha_1 - \beta_4)E]/(\beta_2 - \gamma_1).$$

By using the theoretically predicted signs of the coefficients from the model in equations (1)–(4), we are able to determine the signs of the coefficients of the variables

[16] The supply of private mortgage loans might be influenced by either changes in the value of assets of financial institutions, or changes in the desired portfolio composition of such institutions. The importance of these variables cannot be empirically tested because of data limitations. Nevertheless it is important to consider the impact on the model from their omission. Since most financial institutions held assets that were affected by similar temporal influences (for example, changes in bond prices), what is important are the differences in the desired mix of assets across financial institutions. If farm mortgage lenders desire different portfolio changes —due to factors other than farm earnings—this will affect the supply of private loans to the extent that the mix of financial institutions varied across states. Unfortunately we don't have information for the 1930's on the differing desires to supply farm mortgage credit across private financial institutions. Without this information, we cannot determine the effect on the empirical results from the omission of desired portfolio changes. However, the use of aggregate data on all private creditors reduces the empirical seriousness of the omission of this variable, since some institutions may have been attempting to acquire more farm mortgage loans while others were attempting to shed themselves of such loans. What is of issue here is the variance across states in the mix of farm mortgage lenders. Furthermore, it seems likely that the most important influence on desired portfolio changes was farm earnings which is captured in the model. To this degree, desired portfolio composition is implicitly measured in the earnings variable.

in the reduced-form equations (5) and (6). In the quantity equation: the coefficient $(-\beta_1\gamma_1/(\beta_2-\gamma_1))$ of M is expected to be negative; the coefficient $(-\beta_3\gamma_1/(\beta_2-\gamma_1))$ of S is expected to be positive; the coefficient $((\alpha_1\beta_2\gamma_2-\beta_4\gamma_1)/(\beta_2-\gamma_1))$ of E is expected to be positive.

In the interest rate equation: the coefficient $(-\beta_1/(\beta_2-\gamma_1))$ of M is expected to be positive; the coefficient $(-\beta_3/(\beta_2-\gamma_1))$ of S is expected to be negative; the coefficient $((\gamma_2\alpha_1-\beta_4)/(\beta_2-\gamma_1))$ of E is indeterminate.

We are interested in testing whether the quantity of private farm mortgage loans fell or the interest rates on loans rose relatively more in states that passed moratoria. To make comparisons across states, we need to control for inherently different state credit markets. For example, private lenders in the more agricultural states could be expected to transact a greater quantity of loans. The level of interest rates on loans across states could also be expected to vary with differences in agricultural risk, inter alia, across regions. In addition, actual farm earnings may have deviated from expected farm earnings in some states. Being more or less out of equilibrium might affect the ability and willingness of farm mortgage creditors and debtors to transact new loans.[17] Cross-state differences in credit markets are controlled by estimating the parameters of the following time derivative of the reduced-form equations (5) and (6);

$$(7)\quad \ln(Q_P34/Q_P32)=\delta_0+\delta_1M$$
$$+\delta_2\ln(S_{34}/S_{32})+\delta_3\ln(E_{30-33}/E_{26-29})+e,$$

$$(8)\ \ln(i_{34}/i_{32})=\lambda_0+\lambda_1M+\lambda_2\ln(S_{34}/S_{32})$$
$$+\lambda_3\ln(E_{30-33}/E_{26-29})+e,$$

where Q_P34/Q_P32 = the percentage of all new mortgages in 1934 by private lenders divided by the percentage of all new mortgages in 1932 issued by private lenders;[18]

$M=1$ if the state passed a moratorium in 1932 or 1933 and 0 if it did not;

i_{34}/i_{32} = a weighted average of interest rates on loans issued by private lenders in 1934 divided by a weighted average of interest rates on loans issued by private lenders in 1932;

S_{34}/S_{32} = a weighted average of the average size private farm mortgage loan in 1934 divided by a weighted average of the average size private farm mortgage loan in 1932;

E_{30-33}/E_{26-29} = the average cash receipts from farm marketings and government payments in 1930 to 1933 divided by average cash receipts from farm marketings in 1926 to 1929;[19] and e = stochastic disturbance term.

The coefficients estimated correspond to the reduced-form coefficients solved for in equations (5) and (6). All of the variables used in the regression analysis are the theoretically correct variables as specified in my model except for Q_P. Data do not exist on the number of mortgage loans transacted by private and government creditors. However, my proxy, the percentage of new mortgages by private lenders, must move in the same direction as the quantity of loans.[20]

[17] I am grateful to the referee for pointing out the potential importance of disequilibrium in credit markets across states.

[18] It is not clear whether the value of loans or the number of loans is the preferable proxy for the quantity of loans. In either case, the explanatory variables are expected to have the same directional impact. Ideally I would estimate separate regressions using the value of loans and the number of loans and compare the results. However data do not exist on the value of loans.

[19] Average earnings in 1930–33 as a percentage of average earnings in 1926–29 measures the extent to which earnings deviated from expected trend. Hence it captures both earnings prospects for the future and the condition of credit markets across states.

[20] Q_P and $Q_P/(Q_P+Q_G)$ must move in the same direction. From equation (5) we know that

$$Q_P=[(\beta_2\gamma_0+\beta_2\gamma_2\alpha_0-\gamma_1\beta_0)-\beta_1\gamma_1M-\beta_3\gamma_1S$$
$$+(\alpha_1\beta_2\gamma_2-\beta_4\gamma_1)E]/(\beta_2-\gamma_1).$$

Dividing both sides by Q_P+Q_G gives us the variable I used, $Q_P/(Q_P+Q_G)$. This is equal to $1/(1+(Q_G/Q_P))$. To show that Q_P moves in the same direction as $Q_P/(Q_P+Q_G)$, assume Q_P increases. This implies that $1/(1+(Q_G/Q_P))$ will increase as long as there is either no correlation or a negative correlation between Q_P and

TABLE 4—DESCRIPTIVE STATISTICS

	Mean	Standard Deviation	Minimum	Maximum
Moratorium	.49	.51	0	1
Q_{P34}/Q_{P32}	.40	.17	.09	.79
i_{34}/i_{32}*	.94	.04	.86	1.01
i_{34}/i_{32}**	.94	.04	.83	1.01
$Size_{34}/Size_{32}$*	.95	.14	.73	1.61
$Size_{34}/Size_{32}$**	1.02	.17	.75	1.68
$Earnings_{30-33}$ $/Earnings_{26-29}$	.61	.10	.42	.80
$Earnings_{33}$ $/Earnings_{31}$	.86	.14	.57	1.23

Sources: Moratorium: See Table 1; data on Q_{P34}/Q_{P32}, i_{34}/i_{32}, and $Size_{34}/Size_{32}$ are found in individual state reports issued by the USDA, *Farm Mortgage Recordings*; data for $Earnings_{30-33}/Earnings_{26-29}$, USDA *Cash Receipts From Farming*.

Notes: Moratorium = 1 if the state passed a moratorium, and 0 if it did not.

Q_{P34}/Q_{P32} is the ratio of the percentage of new farm mortgages by private lenders in 1934 to the percentage in 1932. The percentage of new mortgages by private lenders in 1932 and 1934 is the sum of the percentages of new mortgages issued by individuals, national and state banks, insurance companies, mortgage companies, mutual savings banks, and "other" private institutions.

i_{34}/i_{32} is the ratio of a weighted average of interest rates in 1934 charged by each private lending group to interest rates charged in 1932. A single asterisk denotes that the average interest rate constructed is weighted by the percentage of all new private farm mortgages held by each lending group in 1932. A double asterisk denotes that the weights are based on the percentage of new farm mortgages held by each group of creditors in 1934.

$Size_{34}/Size_{32}$ is the ratio of a weighted average of the average size of private farm mortgages in 1934 recorded by the lending groups above to the average size of private farm mortgages in 1932. A single asterisk denotes that the weights are based on the percentage of loans held by various creditors in 1932 while a double asterisk denotes that the weights are based on 1934.

$Earnings_{30-33}/Earnings_{26-29}$ is the ratio of the average of cash receipts from farm marketings and government payments from 1930–33 to the average of cash receipts from farm marketings in 1926–29.

$Earnings_{33}/Earnings_{31}$ is the ratio of cash receipts from farm marketings and government payments in 1933 to cash receipts from farm marketings in 1931.

Descriptive statistics on the data used in the regression analysis, along with the sources of the data, are presented in Table 4. The sample includes observations on states which either passed a moratorium in late 1932 or 1933, or never passed a moratorium. To control for temporal influences on the varying actions taken by creditors but not captured by my explanatory variables, observations on Louisiana, Mississippi, and South Carolina, the states passing moratoria in 1934, are not included in the sample.

The data on the dependent variables were originally collected as part of a nationwide WPA project during 1936 and 1937. For the dependent variables, only actions taken by private lenders are considered. The rationale for this delineation is that governmental lending institutions, especially after 1933, made many of their loans on the basis of "need" rather than expected profits. Hence, a moratorium would not affect the quantity of loans supplied by the government.

It is important to keep in mind that of primary interest to this study is the coefficient of M. Nevertheless, the other variables must be included in the regression analyses to control for their possible influence on Q_P or i. The results reported in Table 5 are consistent with the hypothesis that creditors

Q_G. The only identical causal variable in the Q_G^S or Q_P^S equations (1) and (2) is E, which is hypothesized to influence Q_P^S positively and influence Q_G^S negatively. Therefore it is consistent with the model that Q_P and $Q_P/(Q_P+Q_G)$ move in the same direction.

TABLE 5—*OLS* REGRESSION RESULTS

	Equation: Dependent Variable					
	1: Q_{P34}/Q_{P32}	2: i_{34}/i_{32}	3: Q_{P34}/Q_{P32}	4: i_{34}/i_{32}	5: Q_{P34}/Q_{P32}	6: i_{34}/i_{32}
Constant	−.28	−.04	−.22	−.05	−.80	−.08
	(−1.34)	(−1.79)[b]	(−1.02)	(−2.57)[a]	(−5.45)	(−7.89)[a]
Moratorium	−.29	.02	−.30	.01	−0.37	.006
	(−2.41)[a]	(1.59)[b]	(−2.52)[a]	(.87)	(−2.81)[a]	(.61)
$Size_{34}/Size_{32}$	.53	−.10	.76	−.11	0.28	−0.14
	(1.21)	(−2.36)[a]	(1.58)[b]	(−2.64)[a]	(0.55)	(−3.41)[a]
$Earnings_{30-33}$	1.17	.07	1.19	.05		
$/Earnings_{26-29}$	(2.88)[a]	(1.69)[b]	(3.03)[a]	(1.41)		
$Earnings_{33}$					0.87	−0.03
$/Earnings_{31}$					(0.21)	(−0.93)
R^2	.30	.27	.31	.24	.16	.22

Sources: See Table 4.

Notes: The *t*-statistics are shown in parentheses; $N = 45$. In equations 1 and 2, i_{34}/i_{32} and $Size_{34}/Size_{32}$ are weighted based on the percentage of all new private mortgages in 1934 held by each private lending group. In equations 3, 4, 5, and 6, i_{34}/i_{32} and $Size_{34}/Size_{32}$ are weighted based on the percentage of all new private mortgages in 1932 by each private lending group.

[a] Significant at the 95 percent confidence level.

[b] Significant at the 90 percent confidence level.

took differential action in states passing moratoria. The coefficient on *Moratorium* in equation 1 is consistent with the hypothesis that the percentage of all new loans held by private creditors fell significantly more in states passing moratoria. The coefficient on *Earnings* is also significant with the predicted sign. This result is consistent with the hypothesis that relatively high farm earnings encouraged private creditors and borrowers to transact new farm mortgage loans. The hypothesis that size of mortgage loans influenced the number of mortgage loans is not supported by the regression results. In equation 2, the coefficient on *Moratorium* is consistent with the hypothesis that moratoria legislation induced private creditors to raise interest rates. It is not possible to posit the effect of relative differences in earnings on interest rates because earnings affect both the demand and supply of private mortgage loans. The positive and significant sign on the coefficient of *Earnings* indicates that the outward shift in demand for private loans—because of the government supplying fewer loans when relative earnings increase—exceeds the outward shift in the supply of loans induced by relatively high farm earnings encouraging private creditors to reduce interest rates. Unlike in equation 1, the coefficient on *Size* is consistent with the model's prediction. Where the average size loan increased, perhaps due to the federal government's policy of making small loans and thereby increasing the average size private loan, interest rates fell.

To test the sensitivity of my empirical results, equations 3 and 4 were run weighting i_{34}/i_{32} and $Size_{34}/Size_{32}$ by the percentage of loans held by lenders in 1932. As in equation 1, the coefficient on *Moratorium* in equation 3 is significant indicating that creditors made relatively fewer loans in states passing moratoria. The coefficients on the control variables in equation 3, $Size_{34}/Size_{32}$ and $Earnings_{30-33}/Earnings_{26-29}$, are also significant with the predicted signs. In equation 4, the hypothesis that creditors raised interest rates relatively more in states with moratoria is not consistent with the data. This mixed result on the coefficient on *Moratorium* in equations 2 and 4 may be due to creditors having found it less costly to adjust to the short-run disequilibrium arising from temporary moratoria by simply not granting loans to certain borrowers. In addition, it is reasonable to assume that creditors raised interest rates to some prospective borrowers who at the higher price of credit dropped out of the market. If this conjecture is valid, we would not observe any significant change in interest rates but we would observe a smaller percentage of all new loans issued by private creditors. The stability of the coefficient on *Moratorium* in equations 1 and 3 suggests that it was less costly for creditors to ration the number of loans directly as opposed to indirectly through raising interest rates.

The earnings variable in equations 1, 2, 3, and 4 proxies the extent to which earnings over the several years before a moratorium were below expected earnings. In a sense it measures the degree of disequilibrium. To test my model using a variable capturing more of an equilibrium situation, I estimated the parameters of the model in equations 5 and 6 with $Earnings_{33}/Earnings_{31}$—the ratio of farm earnings in 1933 to farm earnings in 1931. Although the overall fit of the equation is not as good as in the previous equations, the results are similar. The coefficient on *Moratorium* in equation 5 is significant while insignificant in equation 6. The most striking difference in equations 5 and 6 from equations 1, 2, 3, and 4 is that the coefficients on the earnings variable are insignificant. This suggests, perhaps, that the more appropriate characterization of credit markets in the 1930's is disequilibrium.

To determine whether the results are sensitive to aggregating the lending groups, I tested the hypothesis against data on selected lending groups. The results in some instances are more consistent with the hypotheses when the data are disaggregated. When data for insurance companies are used, the coefficient on moratoria in the quantity equation is highly significant. The t-ratio is -3.50 and the R^2 statistic is .59.[21]

The overall results are quite robust, especially in light of the assumption that moratorium legislation in the above regression analysis is identical across all states passing a moratorium. Clearly, the harshness of legislation varied considerably. In Vermont, for example, foreclosure could be postponed for only three months (Central Housing Committee, p. A-26). In general, moratoria in the Midwest were more injurious to creditors. It was not uncommon for moratoria of two years to be enacted and in some instances reenacted for an additional two years (Central Housing Committee, pp. A-3–A-28). I suspect that if we were able to construct a continuous variable according to the severity of the moratorium, the hypotheses tested above would be even more consistent with the data.[22]

III. Conclusion

From 1932 to 1934, twenty-five state legislatures passed laws temporarily preventing the foreclosure of farms by creditors. This paper investigates both the causes and the consequences of the state moratoria on farm foreclosures. Econometric tests show that moratoria were more likely to be enacted in states suffering relatively more farm distress, and less likely to be enacted in states where federal land banks held a relatively greater percentage of the farm mortgage debt. A lesson to be learned from moratoria legislation is that it is not costless. Although the net impact on debtors is impossible to ascertain, it appears as if some debtors gained a temporary reprieve from foreclosure, and some farmers averted foreclosure, but this benefit was at the expense of private creditors and prospective farmers who were precluded from securing credit to purchase a farm because of the increased costs to private creditors.

[21] When data for only national and state banks are used, the substantive results are similar. In the quantity equation, the coefficient on *Moratorium* has the predicted sign and is significant at the 90 percent confidence level.

[22] The reason for my inability to construct a continuous variable is because, not only did the length of the legislated moratoria vary, but the various state judiciaries frequently had discretion in setting rents and other terms that affected the de facto harshness of the legislation.

REFERENCES

Alston, Lee J., "Farm Foreclosures in the United States During the InterWar Period," *Journal of Economic History*, December 1983, *43*, 885–903.

——, "The Role of Financial Institutions

as Sources of Farm Mortgage Credit: 1916–1940," in D. Martin and G. T. Mills, eds., *Dictionary of the History of American Banking*, Westport: Greenwood Press, forthcoming.

Friedman, Lawrence M., *A History of American Law*, New York: Simon and Schuster, 1973.

Hanna, Frank A., *State Income Differentials, 1919–1954*, Durham: Duke University Press, 1959.

Leuchtenburg, William E., *Franklin D. Roosevelt and the New Deal*, New York: Harper and Row, 1963.

Perkins, Van L., *Crisis in Agriculture: The Agricultural Adjustment Administration and the New Deal, 1933*, Berkeley; Los Angeles: University of California Press, 1969.

Saloutos, Theodore and Hicks, John D., *Agricultural Discontent in the Middle West 1900–1939*, Madison: University of Wisconsin Press, 1961.

Shover, John L., *Cornbelt Rebellion: The Farmers' Holiday Association*, Urbana; London: University of Illinois Press, 1965.

Skilton, Robert H., "Developments in Mortgage Law and Practice," *Temple University Law Quarterly*, August 1943, *17*, 326–41.

Woodruff, Archibald M., Jr., *Farm Mortgage Loans of Life Insurance Companies*, New Haven: Yale University Press, 1937.

Home Building and Loan Association v. Blaisdell et al., 290 U.S. 398, 1934.

U.S. Congress, Central Housing Committee, "Digest of State Mortgage Moratorium Legislation and Judicial Interpretation of Same," Appendix No. I, Sub-Committee on Law and Legislation, *Special Report No. 1 on Social and Economic Effects on Existing Foreclosure Procedure and Emergency Moratorium Legislation*, April 2, 1936.

U.S. Department of Agriculture, Bureau of Agricultural Economics, *Cash Receipts From Farming*, Washington: USGPO, January 1946.

______, "Farm Mortgage Credit Facilities in the United States," by Donald C. Horton et al., *Miscellaneous Publication No. 478*, Washington: USGPO, 1942.

______, *Farm Mortgage Recordings*, Washington: USGPO, 1938 and 1939.

______, "The Farm Real Estate Situation, 1933–34," by B. R. Stauber and M. M. Regan, *Circular No. 354*, Washington: USGPO, 1935.

______, "State Measures for the Relief of Agricultural Indebtedness in the United States, 1932 and 1933," compiled by Louise O. Bercaw et al., *Agricultural Economics Bibliography No. 45*, Washington: USGPO, 1933.

______, "State Measures for the Relief of Agricultural Indebtedness in the United States, 1933 and 1934," compiled by Louise O. Bercaw et al., *Agricultural Economics Bibliography No. 53*, Washington: USGPO, 1934.

U.S. Department of Commerce, Bureau of the Census, *Fifteenth Census of the U.S., 1930: Agriculture*, Vol. II, Washington: USGPO, 1932.

______, *Statistical Abstract of the United States —1932*, Washington: USGPO, 1934.

The Economics of Performing Shakespeare

By JAMES H. GAPINSKI*

> The play's the thing
> Wherein I'll catch the conscience of the King.
>
> *Hamlet*

With an origin dating back to David Garrick's Shakespeare Jubilee of 1769 held at the Bard's birthplace Stratford-upon-Avon, the Royal Shakespeare Company (*RSC*) made a formal debut in 1961, the year of its christening by Her Majesty the Queen. Besides *As You Like It*, its early repertoire included *King Lear*, *Macbeth*, and *Richard III* variously starring such notables as Peggy Ashcroft, Vanessa Redgrave, Diana Rigg, and Paul Scofield and staged at the Shakespeare Memorial Theatre in Stratford or the Aldwych Theatre in London. From these two proscenia, the former also known as the Royal Shakespeare Theatre, the *RSC* played to more than a half million people annually until 1982, when it moved from the modest and cramped quarters of the Aldwych to the lush and vast surroundings of London's newly completed Barbican Centre.

A nonprofit group, the Company is one of four such performing arts organizations in Britain to be designated as "national." Its designation seems to be apt because historically the Company has endeavored to reach beyond the artistic and physical limits of its twin headquarters. Since the mid-1970's, for instance, it has been offering experimental theater from The Other Place in Stratford and from The Warehouse—now relocated to The Pit—in London. It regularly stages productions in Newcastle-upon-Tyne, and it frequently broadcasts its efforts over television. It also tours internationally, its latest overseas achievements including the New York engagement of the marathon *Life and Adventures of Nicholas Nickleby* and the Broadway run of *Good.*

From all of its artistic activities, Shakespearean and otherwise, the Company succeeded in keeping its earned income ahead of inflation. According to Table 1, real earned income rose from £1.40 million in financial year 1965–66 (April 1965 to March 1966) to £1.84 million in 1979–80, and throughout that decade and a half, it grew at an average yearly rate of 1.34 percent. On the opposite side of the accounting ledger, however, stand the Company's expenses, which have continually outstripped earned income by a wide margin as the table reveals. That discrepancy, called the income gap and acknowledged as a hallmark of the nonprofit lively arts operating on either side of the Atlantic (William Baumol and William Bowen, 1966, pp. 147–50; 474), has been countered by patronage originating principally from the Art Council of Great Britain. Over time patronage has become an increasingly important component of *RSC* income, but, despite its magnitude, large year-end deficits have proven to be the rule.

Given its long tradition, its international presence, and its budgetary magnitude, the *RSC* must be recognized as a major cultural force. Nonetheless, very little has been done to determine the economic underpinnings of its operations. The Company uses labor and capital to generate output; yet its production function has escaped study. It sells output in the marketplace; yet its demand curve has remained hidden. It relies heavily on subsidy; yet the consequences of that support have not been quantified. Do the laws of production apply to the *RSC*? Do the laws of demand apply? Does the benefit from the

*Florida State University, Tallahassee, FL 32306. I thank Gordon Andrews of the University of Birmingham, Margaret D'Andrade of the Royal Shakespeare Company, George Darroch of the Arts Council of Great Britain, Trevor Gambling of UB, I. M. Hutchings of the Department of Inland Revenue, and Graham Jenkinson of the Department of Trade for help in developing the data file. I also thank Edward McIntyre of FSU and, once again, Gordon Andrews for help in making some sense of it. Thanks likewise go to Frederick Bell of FSU, and to an anonymous referee for comments on the initial draft of this paper. The responsibility for views and errors herein rests, I regret, solely with me.

TABLE 1—A FINANCIAL RECORD OF THE *RSC*, DEFLATED SERIES[a]

Financial Year	Total Earned Income	Total Patronage	Total Income[b]	Total Expenses	Total Deficit[c]	Patronage Ratio[d]
1965–66	1,404,477	216,895	1,621,372	1,674,128	−52,756	.134
1966–67	1,356,114	335,165	1,691,279	1,692,255	− 976	.198
1967–68	1,449,185	440,860	1,890,045	1,960,015	−69,970	.233
1968–69	1,546,112	467,719	2,013,831	2,341,990	−328,159	.232
1969–70	1,593,052	458,236	2,051,288	2,189,481	−138,193	.223
1970–71	1,731,051	507,541	2,238,592	2,226,908	11,684	.227
1971–72	1,849,664	500,992	2,350,656	2,355,707	− 5,051	.213
1972–73	1,697,566	663,772	2,361,338	2,368,079	− 6,741	.281
1973–74	1,846,029	572,569	2,418,598	2,440,975	−22,377	.237
1974–75	1,541,884	823,263	2,365,147	2,373,406	− 8,259	.348
1975–76	1,538,997	754,053	2,293,050	2,292,295	755	.329
1976–77	1,603,770	1,017,274	2,621,044	2,621,044	0	.388
1977–78	1,645,473	1,084,964	2,730,437	2,770,052	−39,615	.397
1978–79	1,715,715	1,275,715	2,991,430	3,062,538	−71,108	.426
1979–80	1,843,393	1,315,246	3,158,639	3,193,915	−35,276	.416

Sources: Nominal figures come from *RSC* annual reports issued by the Council of the Royal Shakespeare Theatre. The deflator is the Retail Price Index compiled by the Department of Employment. Based at 1.00 for calendar year 1975, it is recast in financial years.

[a]Except for the patronage ratio, all entries are denominated in units of pounds.

[b]Total income = total earned income + total patronage.

[c]Total deficit = total income − total expenses.

[d]Patronage ratio = total patronage/total income.

subsidy outweigh its cost? These issues form the subjects of the present paper.

Section I is devoted to the production question. Section II takes up the demand matter while Section III makes the patronage inquiry. Section IV collects conclusions and offers comments. Because of severe data limitations involving *RSC* activities beyond the Memorial Theatre—hereafter referred to as Stratford—and the Aldwych, all analyses are restricted to those centers.

I. The *RSC* Production Structure

Work on the production structures of U.S. nonprofit lively arts (see my 1980 article) adopted the transcendental production function, whose special case of the Cobb-Douglas may be written as

$$Q_{it}=\beta_1 e^{\beta_2 Z_{it}} e^{\beta_3 t} L_{it}^{\beta_4} K_{it}^{\beta_5}, \tag{1}$$

where Q_{it} represents the quantity of cultural experiences generated by organizational unit i at time t, while L_{it} denotes the quantity of labor, and K_{it} the quantity of capital used by i at t. The subscripts $i=A$ and $i=S$ indicate Aldwych and Stratford, respectively. The Z_{it} signifies a dichotomous shift variable introduced to account for structural differences across the two theaters; in particular, $Z_{At}=0$ and $Z_{St}=1$ for all t. Technical progress is posited to be disembodied and to be occurring at the rate β_3. Moreover, by virtue of the Cobb-Douglas formulation, factor proportions are taken to be variable to an extent specified by a unitary elasticity of substitution. It is true that a *Hamlet* is not quite the same without poor Yorick's skull, and that a *Macbeth* loses in a translation that omits the witches' cauldron. But, although a given play may require fixed proportions, a repertory entity does have the option of altering proportions during a season by substituting among the plays to be staged. Hence a unitary elasticity need not be at odds with the nature of the production process.[1]

[1] A unitary elasticity in the two-factor case actually conforms to estimates obtained for theater in the three-factor instance. The substitution elasticities reported in my earlier article (p. 584) for the labor and capital combinations lie on either side of one and average .94, a value close to the Cobb-Douglas mark.

The data required to estimate equation (1) came from several sources. Cultural experiences, interpreted as paid attendance in units, were counted by the Arts Council and were reported by financial year from 1965–66 to 1980–81. Although a few holes appeared in the series for each theater, they were quickly filled with the aid of information supplied directly by the *RSC*. Labor and capital numbers were prepared from nominal cost figures tabulated by Trevor Gambling and Gordon Andrews (1982). Examining the ten years 1968–69 to 1977–78, and thereby narrowing the present data set to the same period, Gambling-Andrews identified eight divisions of the *RSC*: players, senior management, administration, production, stage operations, theater operations, publicity, and miscellaneous. Except for the senior management, administration, and miscellaneous categories, *RSC* costs in the various divisions were apportioned between the Aldwych and the Stratford by Gambling-Andrews to gauge the activity originating from each center individually.

In quantifying the labor input for present purposes, it seemed prudent to regard at the outset *RSC* personnel as representing two different types of labor: artists and staff. Artists consisted of the players and the senior management inasmuch as the latter group included th‿ artistic director, directors, and designers. Although senior management costs had not been apportioned by Gambling-Andrews, that task was easily accomplished by appealing to the distribution of player remuneration between centers. Staff included everyone else: from carpenters to painters, from dressers to showmen electrics, and from program sellers to secretaries. This component required the apportioning of administrative and miscellaneous costs, and the procedure followed rivaled the one used for senior management: it relied on the distribution of nonartist remuneration.[2]

What eventually emerged for each theater were the expenditures on artists and staff by financial year. These expenses, expressed in units of pounds, were converted into units of man-hours through division by the average hourly earnings in manufacturing and other industries recorded in *Regional Trends* of the Central Statistical Office (CSO).[3] To capture the regional variation of British wages, the hourly rates pertinent to the South East were adopted for the Aldwych, while those applicable to the West Midlands were selected for the Stratford. Drawn in units of pounds, the wage rates were translated into a financial-year basis before the calculations were made. Artist and staff man-hours were then summed to give the labor variable for each theater.

The capital variable fashioned for each center was a portmanteau measure including those expenses in the Gambling-Andrews data not classifiable as labor costs. Rent, utilities, taxes, insurance, repairs, depreciation, royalties, postage, and materials among other items found their way into the capital category consonant with the notion of service flow.[4] Each theater's capital expenses, in units of pounds, were then divided by a U.K. capital-price index (*KPI*) constructed as an average of the indices relevant to the costs of housing (including rent, taxes, utilities, and repairs), clothing, furniture, radio and electrical goods, books and newspapers, and motor cars. The components, derived from nominal and real consumer expenditure series appearing in the CSO *Annual Abstract of Statistics*, had a base value of 1.00 in calendar year 1975; they were reexpressed in terms of the financial year.

Estimating equation (1) in logarithms by ordinary least squares (*OLS*) on the 20 observations comprising the pooled data file for

[2] To determine the sensitivity of the estimated production function to the apportionment rule chosen, a second rule was tried. Specifically, the costs of senior management were allocated according to the number of plays staged in each theater, and the costs of administration and miscellany were divided according to the number of performances. The differences in rules led to negligible differences in the estimated function.

[3] Instead of the average hourly earnings in manufacturing and other industries, those in theater or, more generally, in service industries would have been preferred as the conversion factor. They were not available, however.

[4] On the matter of rent, Stratford posed a problem because, unlike the Aldwych, it paid none, and therefore no rental costs were charged against it in the Gambling-Andrews report. To estimate that charge, the Aldwych rental figures were adjusted to allow for Stratford's greater seating capacity, and the average of the derived figures was treated as a fixed annual mortgage payment.

the Aldwych and the Stratford from 1968–69 to 1977–78 disclosed a consistent pattern: $\hat{b}_1$ and $\hat{\beta}_3$, with b_1 symbolizing $\ln \beta_1$ and with the circumflexes signifying estimates of the respective parameters, were insignificant,[5] regardless of whether the constant and the time variable were retained in the regression separately or jointly.[6] The estimated rate of technical progress, $\hat{\beta}_3$, was especially weak. Its showing, which agrees with the argument by Baumol-Bowen (pp. 162–67) and with the finding by myself (pp. 582–83), ostensibly says that not much had been done to organizational patterns at the Aldwych and the Stratford during the years under consideration.

This lackluster performance of the constant and the time variable recommended that equation (1) be rerun without them, the result being posted in Table 2. Because the data underlying the fitted equation have both time-series and cross-section dimensions, checking for autocorrelation and heteroscedasticity seems to be obligatory.[7]

TABLE 2—REGRESSION RESULTS FOR THE PRODUCTION AND DEMAND FUNCTIONS[a]

Descriptor	Production Function	Demand Function
$\hat{\beta}_2$	.39341 (7.706)	
$\hat{\beta}_4$	.62405 (5.128)	
$\hat{\beta}_5$	.33191 (2.625)	
$\hat{\xi}$		$.417 \times 10^{-2}$ (10.250)
$\hat{\xi}_2$		$-.199 \times 10^{-2}$ (−2.515)
$\hat{\xi}_4$		$.819 \times 10^{-5}$ (5.743)
$\bar{R}^2$	.87	.96
F	41.32	199.02
$\hat{\rho}$		−.40
D-W	1.95	2.19
V	.279	−.553
Sample Size	20	28

[a] t-values are shown in parentheses.

Table 2 addresses the first issue. There the Durbin-Watson statistic (D-W) suggests the absence of positive and of negative autocorrelation at the 1 percent level, and this suggestion receives confirmation from a two-sided test of the residuals' sign pattern (N. R. Draper and H. Smith, 1966, pp. 95–97). That test statistic V, a normal deviate, falls well below the critical value at the 2 percent level thereby pointing to acceptance of the null hypothesis of random sign arrangement.

Checking for heteroscedasticity follows the lines sketched by J. Benus et al. (1976, p. 133). The residuals are first separated into three different configurations based on the values of the dependent variable, $\ln Q_{it}(=q)$. One configuration consists of three partitions: those residuals corresponding to the 7 lowest q, those corresponding to the 7 middle q, and those corresponding to the 6 highest q. Another configuration follows a 7–6–7 split. A third entails two partitions; namely, the residuals associated with the 10 smallest q and those linked to the 10 largest. This 10–10 configuration amounts to a grouping by theater: the residuals for the Aldwych vs. the Stratford, respectively. For each configuration, variances of the residuals in the partitions are calculated and compared pairwise by an F test (John Freund, 1962, pp. 271–74) to determine if they differ significantly. Of the seven two-sided tests conducted across all configurations at the 2 percent level, none reject the null hypothesis of variance homogeneity.

The estimated Cobb-Douglas appears to be free from both autocorrelation and heteroscedasticity. It also performs well in other respects. Its $\bar{R}^2$ and regression F are quite respectable, and each of its coefficients is significantly different from zero. Its shift parameter assumes a positive value confirming the impression that the Stratford generates more output than does the Aldwych, *ceteris paribus*. Both output elasticities fall in the usual range, but labor's is almost twice

[5] Two-sided t-tests proceed at the 5 percent level; one-sided tests occur at the 2.5 percent level.

[6] Preliminary regression analysis of the data went beyond the confines of specification (1). Two experiments, rather than invoking a single labor variable, ran the pair artists and staff in logarithms and then, to mirror a transcendental function, ran the pair with only artists in logarithms. Another experiment allowed the output elasticities as well as the intercept to vary across theaters. All three exercises gave uninspiring results.

[7] The two autocorrelation tests and the 10–10 heteroscedasticity check described anon were applied to the six preliminary regressions. In every case they intimated that neither anomaly held.

capital's, supporting the intuitive notion that art is primarily the artist's—or, more broadly, labor's—medium. The sum of the elasticities indicates decreasing returns to scale: it is significantly less than one.

Customary laws of production evidently hold for the *RSC*, and in this additional respect the Company resembles its nonprofit cousins on the far side of the Atlantic (see my earlier article, pp. 582–86).

II. Demand Analysis

Cultural experiences, which cannot be inventoried, are sold at the moment of their production. Under the standard precepts the quantity Q_{ijt} of experiences demanded at time t by individual j from theater i may be represented as

$$(2)\qquad Q_{ijt}=\xi_1+\xi_2P_{it}+\xi_3U_{it}+\xi_4Y_{jt},$$

where P_{it} denotes the real price of an experience from i at t, U_{it} signifies the real price at t of an experience that substitutes for one from i, and Y_{jt} represents the individual's real income at t. Again i equals A for Aldwych and S for Stratford. Summing Q_{ijt} for each theater across N_t individuals and consolidating the separate aggregations for the Aldwych and the Stratford into a single expression yield

$$(3)\qquad Q_{it}/N_t=\xi_1+\zeta Z_{it}+\xi_2P_{it}+\xi_3U_{it}+\xi_4Y_t/N_t,$$

where $Q_{it}=\Sigma_jQ_{ijt}$ and $Y_t=\Sigma_jY_{jt}$. The Z_{it} continues as a dichotomous shift variable satisfying the condition $Z_{At}=0$ and $Z_{St}=1$ for all t.[8]

As noted in Section I, the data on quantity came from the Arts Council and the *RSC*, and covered the financial years 1965–66 to 1980–81. Nominal own-price. derived by dividing attendance into box-office receipts inclusive of the value-added tax, originated from the same sources; it was expressed in units of pounds. Nominal substitute price took three different forms. In one version it was the price index for entertainment and recreational services, while in a second it was the average nominal price of admission to cinemas in Great Britain. The third version resembled the second except that it captured regional price variation; in particular, it consisted of cinema prices pertinent to the South East (for $i=A$) and the West Midlands (for $i=S$).[9] The *Annual Abstract of Statistics* served as authority for all three. The entertainment price index was based at 1.00 in calendar year 1975, and both cinema prices were denominated in unit pounds. Nominal income, defined as after-tax total personal income in Great Britain and calibrated in unit pounds, was drawn mostly from *Regional Trends* although some entries had to be obtained directly from the Department of Inland Revenue. To deflate the nominal measures, the Retail Price Index (*RPI*) prepared by the Department of Employment was selected. Like the entertainment price index, the *RPI* had a base of 1.00 in calendar year 1975. Great Britain population counts, in units, included all age groups, their source being *Regional Trends*. Series which were not supplied in a financial-year format were converted to that format. Due to the unavailability of some numbers beyond 1979–80, it was necessary to truncate all series at that point, and consequently the data file closed spanning the fifteen years from 1965–66 to 1979–80.

Equation (3) was fitted by *OLS* to the pool of 30 observations for the Aldwych and the Stratford. Run sequentially under the three definitions of substitute price, it left distinct

[8] Two elaborations of equation (3) were specified. One allowed for partial adjustment while another allowed the coefficients of own-price, substitute price, and per capita income to join the intercept in varying across theaters. Both elaborations failed at the regression stage.

[9] Granted that such substitute prices are not without precedent in studying the demand for nonprofit theater (Susan Touchstone, 1980, pp. 36–37), they nevertheless leave something to be desired here because theatergoers may not regard golf or "Superman III" as a replacement for a live performance of *Richard III*. They may instead choose among the lively arts opting for symphony, opera, dance, or other theater rather than the *RSC*. Price series for several performing arts organizations that likely compete with the Company were constructed from Arts Council data, but they did not contain enough observations to warrant retention in the regression work.

traces of negative autocorrelation and cautioned about the significance of the intercept and the substitute-price coefficient. To address the autocorrelation problem and at the same time to pursue the significance issue, variations of equation (3) were estimated over a grid of values for the first-order Markov coefficient ρ. The variations treated emerged by deleting the intercept and the substitute price individually and together, and given three price definitions, eight expressions were produced. At a cost of one observation per theater, each expression was transformed by the rule $T_t - \rho T_{t-1}$ for any variable T, and it then was estimated under twelve alternative ρ values ranging from $-.10$ to $-.90$. The preferred fit was the one that minimized the residual sum of squares over the dozen, its ρ being designated $\hat{\rho}$. All eight preferred renditions that obtained—one for each of the eight variations of equation (3)—corresponded to internal ρ, and all were free from autocorrelation and heteroscedasticity. However, seven of the eight could be dismissed because of intercept or substitute-price insignificance. Only one rendering remained; it is reported in Table 2.[10]

As the table indicates, the demand equation has acceptable $\bar{R}^2$ and regression F values. In keeping with a property discovered for the production function, the demand shift parameter is positive. The own-price coefficient is negative while the income coefficient is positive. All coefficients differ significantly from zero. Calculated at the means for the period 1965–66 to 1979–80, the price and income elasticities amount to $-.657$ and 1.327, respectively. Apparently demand is price inelastic, and an *RSC* cultural experience is a luxury item. These results accord with intuition.

III. Subsidy: Its Effects, Benefit, and Cost

Information contained in the production and demand functions listed in Table 2 permits an inquiry into the effects of patronage received by the *RSC*. It also enables a quantification of the benefit arising from the subsidy and a comparison of that benefit with its cost.

Appealing to the two equations and to their underlying data, Table 3 presents profiles of the Aldwych and Stratford operations in actual practice. An overbar indicates a mean value calculated from the common period of the production and demand analyses, 1968–69 to 1977–78. A tilde signifies a number calculated from the means. For example, $\tilde{Y}=\bar{X}/\bar{R}$, $\tilde{W}=\bar{W}/\bar{R}$, and $\tilde{M}=\bar{M}/\bar{R}$, where X and R notate nominal income and the *RPI*, respectively, and where W and M represent the nominal wage rate and the *KPI*, respectively. All denominations applicable earlier continue. Again, $Z=0$ for Aldwych, and $Z=1$ for Stratford.

Corroborating some evidence offered in Sections I and II, Table 3 acknowledges the Aldwych to be the smaller of the two outfits under actual conditions. It is smaller in labor and capital and thus in output, its output being roughly half that of Stratford's. Both theaters, however, use labor intensively as the capital-labor ratios reveal. Both also show the marginal product of labor Λ to exceed the marginal product of capital Ω, a result echoing the sentiment, gathered from the output elasticities, that art is mainly labor's medium. Real price, which is lower at the Aldwych, combines with quantity to determine real total revenue Φ, Aldwych's being about half of Stratford's. For each theater that revenue is insufficient to prevent real profit Π from assuming a large negative value: the Aldwych and the Stratford operate with sizable losses!

How do these profiles differ from those that would occur if the centers conducted business on a profit-maximizing basis? With $\tilde{W}$ and $\tilde{M}$ taken as predetermined, maximizing real profit separately at the Aldwych and the Stratford subject to the production and demand functions in Table 2 yields eight equations in eight unknowns: for each theater, real ticket price and the quantities of labor, capital, and cultural experiences. The results of such optimization appear in Table 3. The values $\bar{K}$ and $\bar{L}$ are now interpreted as

[10] Examination for autocorrelation and heteroscedasticity during the search process involved the two autocorrelation tests and the 10–10, now 14–14, heteroscedasticity check described previously. The demand function in Table 2 was also subjected to 9–9–10 and 9–10–9 heteroscedasticity inquiries and passed both.

TABLE 3—OPERATIONS PROFILES OF THE *RSC* UNDER ACTUAL AND PROFIT-MAXIMIZING CONDITIONS

	Aldwych Theatre		Stratford Theatre	
Variable	Actual Conditions	Profit Maximization	Actual Conditions	Profit Maximization
$\bar{L}$	491,040	29,592	580,280	234,640
$\bar{K}$	299,680	20,053	373,870	162,402
$\bar{K}/\bar{L}$	.610	.678	.644	.692
$\tilde{Q}=e^{\hat{\beta}_2 Z}\bar{L}^{\hat{\beta}_4}\bar{K}^{\hat{\beta}_5}$	234,045	16,528	414,281	178,536
$\tilde{\Lambda}=\hat{\beta}_4\tilde{Q}/\bar{L}$	.297	.349	.446	.475
$\tilde{\Omega}=\hat{\beta}_5\tilde{Q}/\bar{K}$	.259	.274	.368	.365
$\tilde{P}=(\tilde{Q}-\hat{\xi}Z\bar{N}-\hat{\xi}_4\tilde{Y})/(\hat{\xi}_2\bar{N})$	1.77	3.78	2.19	4.38
$\tilde{\Phi}=\tilde{P}\tilde{Q}$	414,260	62,532	907,275	781,326
$\tilde{\Pi}=\tilde{\Phi}-\tilde{W}\bar{L}-\tilde{M}\bar{K}$	−504,571	5,181	−213,843	316,848

the optimal input levels and not as the sample means. The other endogenous variables are reinterpreted accordingly. Relative to a profit-maximizing stance, the *RSC* in actual practice overproduces experiences at both centers. Aldwych overproduces by more than a factor of 14; Stratford, by more than a factor of 2. The labor and capital inputs are used to excess at both. Given the conventional shape of the demand curve facing each theater, real ticket prices fall below the profit-maximizing levels. Optimization would require prices to more than double at the Aldwych and to exactly double at the Stratford. Nevertheless, obeying the maximization principles would turn a box office profit. At the Aldwych, real profit would rise from −£505,000 to £5,000 while at the Stratford it would soar from −£214,000 to £317,000.[11]

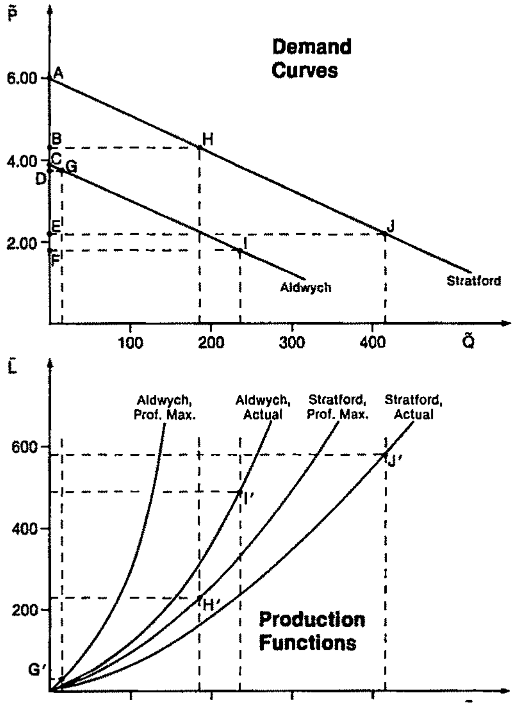


FIGURE 1

Patronage, then, leads to lower prices, to increased cultural experiences, and to additional inputs. More people are exposed to Shakespearean theater, and more artists, carpenters, dressers, and secretaries are employed than would be true without the subsidy. Figure 1 depicts the comparison. There, as in Table 3, price is denominated in unit pounds. Cultural experiences, however, are expressed in thousands of units, and labor is set in thousands of man-hours. Points *I* and *J* identify the actual demand positions of the Aldwych and the Stratford, respectively, while points *I′* and *J′* site the corresponding production positions; *G* and *H* are the profit-maximizing demand points and *G′* and *H′* are the optimum production points.

The subsidy going to the *RSC* has the expected desirable effects. But are those

[11] It may bear repeating that the large differences noted come from comparing actual values of the *RSC* with target values defined by the optimum for a hypothetical profit maximizer. They neither say nor require that the optimality conditions, which are cast in terms of marginals, hold for large changes in the relevant variables.

effects justified in view of their cost? Does the benefit of the gift exceed its cost?[12] To answer that question the benefit must be quantified, and one method for doing so treats the gift's benefit as the resulting gain in consumer's surplus.[13] If the *RSC* were an unsubsidized profit maximizer, it would locate the Aldwych's price and quantity at *G* in Figure 1, creating a real consumer's surplus represented by the triangle *CGD*: £1,266. Locating instead at point *I*, the Aldwych engenders a real surplus of *CIF* or £253,841. For the Stratford, profit maximization would yield the surplus triangle *AHB* as opposed to the actual triangle of *AJE*, £147,712 vs. £795,341, respectively. Hence the *RSC* subsidy increases real consumer's surplus by £900,204. Mean nominal patronage for the period 1968–69 to 1977–78 registers £610,616, which becomes £761,937 in real terms. It follows that *RSC* patronage has a benefit-cost ratio of 1.18: each £1.00 of subsidy returns £1.18 in benefit. The subsidy more than "pays" for itself, and by the customary argument it is justified.[14]

IV. Conclusions and Comments

Economic laws hold at the Royal Shakespeare Company. The relationship between inputs and output obeys a well-behaved, standard production format; the linkages among quantity, price, and income observe established demand principles; and the subsidy received affects activity in the anticipated way. By its economics the Company closely resembles nonprofit lively theater in the United States.[15] It has an output elasticity for labor that exceeds the output elasticity for capital and a marginal product of labor that exceeds the marginal product of capital. It evidences decreasing returns to scale, and it heeds an elasticity of substitution that seems to at least approximate unity. It appears to have missed the advance of technology. It faces a demand schedule that is price inelastic and income dependent, but one that is perhaps unaffected by pricing events in other leisure pursuits. It uses labor and capital to excess from a profit-maximizing standpoint, it underprices the cultural experience that it offers, and it occasions huge losses financed largely through patronage. By virtue of these similarities between the *RSC* and nonprofit theater in the United States, the findings reported here have a scope that goes beyond the boundaries of a single organization.

Nonprofit performing arts lend themselves to formal and traditional economic modeling. This trait should be consoling to public authorities because their decisions on the arts have economic consequences—consequences for price, quantity, employment, and capacity. Arts models have much to say about what those consequences might be, and thus they represent a useful tool for policy formulation. They may be especially useful at a time when public officials seek ways to cut a government deficit. The arts might be viewed in the legislative and executive branches as clear candidates for austerity because they seem to provide a luxury item for the few. The view from this paper is not at all that clear, inasmuch as decreased funding of the arts would reduce employment and output, would raise price, would destroy surplus, and would compromise a government program that has merit on economic grounds. The issue is hardly one-dimensional.

Research papers often close with an admission that much more must be done. This

[12] Benefit-cost analysis is popular for evaluating the provision of public goods, and inasmuch as an *RSC* cultural experience can be construed as being at least partly public (Baumol-Bowen, pp. 380–86), the application of such analysis in the present circumstance seems to be appropriate.

[13] Support for using consumer's surplus to gauge the benefit in the applied context can be deduced from, say, Arnold Harberger (1971).

[14] If the patronage figure were calculated as the mean of the reals reported in Table 1 rather than as the quotient of the means of nominal patronage and *RPI*, it would read £685,038, and the benefit-cost ratio would be 1.31, thereby strengthening the case for the subsidy. That case would be further strengthened if the subsidy number entering the arithmetic were the one pertinent to Aldwych and Stratford activities alone. Isolating Aldwych and Stratford patronage from total patronage was not attempted because the information on patronage allocation proved to be incomplete.

[15] The U.S. experience is described in Baumol and Bowen (pp. 147–50; 162–67), my article (pp. 582–84), and Touchstone (pp. 36–39).

paper closes with a recognition that not much more can be done—not without new, comprehensive data. A decade ago the Ford Foundation (1974) made available a rich source covering the U.S. nonprofit lively arts, but that material has become tired and worn. Nothing as ambitious has been undertaken since, and nothing of its kind has ever been undertaken in the United Kingdom. Instead, research has had to proceed on the basis of individual efforts at collecting small samples, the sum of which fails to yield anything even remotely related to a consistent whole. Development of a consistent whole likely requires the support of agencies such as the Arts Council and the National Endowment for the Arts. Those entities are principal ones to decide if an appreciable furthering of cultural economics is to be or not to be.

REFERENCES

Baumol, William J. and Bowen, William G., *Performing Arts—The Economic Dilemma,* New York: Twentieth Century Fund, 1966.

Benus, J., Kmenta, J., and Shapiro, H., "The Dynamics of Household Budget Allocation to Food Expenditures," *Review of Economics and Statistics,* May 1976, *58,* 129–38.

Draper, N. R. and Smith, H., *Applied Regression Analysis,* New York: Wiley & Sons, 1966.

Freund, John E., *Mathematical Statistics,* Englewood Cliffs: Prentice-Hall, 1962.

Gambling, Trevor and Andrews, Gordon, "An Analysis of the Personnel Costs of a Major Theatrical Company: 1968–78," unpublished paper, University of Birmingham, England, 1982.

Gapinski, James H., "The Production of Culture," *Review of Economics and Statistics,* November 1980, *62,* 578–86.

Harberger, Arnold C., "Three Basic Postulates for Applied Welfare Economics: An Interpretive Essay," *Journal of Economic Literature,* September 1971, *9,* 785–97.

Touchstone, Susan Kathleen, "The Effects of Contributions on Price and Attendance in the Lively Arts," *Journal of Cultural Economics,* June 1980, *4,* 33–46.

Central Statistical Office, *Annual Abstract of Statistics,* London: HMSO, various years.

______, *Regional Trends* (formerly *Regional Statistics* and *Abstract of Regional Statistics*), London: HMSO, various years.

Council of the Royal Shakespeare Theatre, *Report of the Council,* Stratford-upon-Avon, various years.

Ford Foundation, *The Finances of the Performing Arts,* Vol. 1, New York: The Ford Foundation, 1974.

Econometric Policy Evaluation: Note

By Thomas F. Cooley, Stephen F. LeRoy, and Neil Raymon*

In Robert Lucas's (1976) representation of the received method of econometric policy evaluation, a government policy is characterized by a sequence of values of a policy variable x_t. According to the Keynesian tradition, the effect of a given policy on such endogenous variables as *GNP* is determined by solving an econometric model

$$y_{t+1} = F(y_t, x_t, \theta, \varepsilon_t)$$

for successive values of the endogenous variables y_t, with the x_t treated as deterministic forcing variables. Here θ is a parameter vector and the ε_t are random shocks. Lucas correctly observed that such a formulation is inconsistent with a view of agents as optimizers: except in special cases in which the future is irrelevant to present decisions, it makes no sense to think of agents as optimizing if they know that their budget constraints are liable to shift arbitrarily (i.e., in a way which is not characterized probabilistically) as government policy changes. Lucas was led to augment the foregoing equation by adding to the system a government policy function

$$x_t = G(y_t, \lambda, \eta_t).$$

Since the η_t are random variables, a probability distribution is induced on the x_t. Alternative response functions may be viewed as indexed by the parameter vector λ, and policy may be subject to random shifts, indexed by η_t, due perhaps to the vagaries of the political process.

For Lucas, a policy evaluation exercise is conducted by comparing the probability distribution of the y_t for different hypothetical values of λ. This procedure presumes that agents act as if they are certain what policy rule is in force and, further, act as if they are certain that the rule will be maintained into the indefinite future. These are very severe limitations. First, under rational expectations agents will attach very high probability to the event that the current regime will prevail into the indefinite future only if regime changes of the type under consideration are in fact very rare. Thus Lucas's framework is applicable only to a small minority of policy changes. This point was noted by Christopher Sims (1982, pp. 118 ff.). Second, even if the regime change under discussion is in fact a rare event, the analysis still is relevant only after agents have convinced themselves with certainty that a new regime is in place.

There are several possible responses to these limitations on the set of policy changes analyzable in the way Lucas outlined:

Accept these limitations as inherent in the nature of policy analysis. In 1980, for example, Lucas explicitly rejected the idea that economists should undertake to evaluate government policy in specific historical episodes such as, in this case, the 1974–75 recession.[1] This stance is consonant with Lucas's formal policy evaluation procedure, as that is incapable of generating a compari-

*Cooley and LeRoy: Department of Economics, University of California, Santa Barbara, CA 93106; Raymon: Department of Economics, University of Missouri, Columbia, MO 65211.

[1] Here it is important to observe that Lucas's point was not simply that policy actions conducted in 1974–75 were undertaken in response to events occurring before 1974, and had consequences after 1975, implying that the interval over which the analysis is conducted should be lengthened. That point was in fact made by William Poole (1980) at the same conference, and is entirely uncontroversial. Lucas's rejection of policy evaluation in specific historical episodes was more sweeping:

> What advice, then, do advocates of rules have to offer with respect to the policy decisions before us *right now*? This question does have a practical, men-of-affairs ring to it, but to my ears the ring is entirely false. It is a king-for-a-day question which has no real-world counterpart in the decision problems actually faced by economic advisors.... Economists who pose this "What is to be done, today?" question as though it were somehow the acid test of economic competence are culture-bound (or institution-bound) to an extent they are probably not aware of. They are accepting as *given* the entirely unproved hypothesis that the fine-tuning exercise called for by the Employment Act [of 1946] is a desirable and feasible one. [1980, p. 208]

son between two proposed policy sequences evolving out of a common past (since policy is identified with λ, a constant).

Lucas's attitude owes also to another consideration. He has expressed the view that it makes no sense to think of the government as conducting one of several possible policies while at the same time assuming that agents remain certain about the policy rule in effect. Under changing policy rules, Lucas finds that the assumption of rational expectations becomes implausible.[2] For example, Lucas and Thomas Sargent wrote that in this kind of environment "it is...[hard] to imagine that agents can successfully figure out the constraints that they face" (1981, p. *xxxvii*). But if the assumption of rational choice is inapplicable, then there is no hope that we can understand or predict the effects of policy changes.

Expand the analysis to allow for learning. John Taylor (1975) and others have responded to the criticism that agents are assumed to know with certainty the value of λ by instead assigning to agents a nondegenerate prior on λ at the time of the regime change. Then it is assumed that agents update their subjective distributions on λ according to a Bayesian learning process. This analysis disposes effectively of the problem of accounting for how agents behave before becoming certain of the new value of λ, at least if we ignore the question of where their priors come from. However, in these analyses it is still assumed that agents are certain that the regime parameter does not change, even though they do not know its value with certainty. Thus again either the analysis is limited to those very few regime changes which can be regarded by agents as virtually permanent once they occur, or agents are being modeled as having nonrational expectations about the likelihood of future regime changes.

Use nonstructural methods for policy evaluation. Sims has responded to these difficulties by recommending that policy evaluation be conducted using nonstructural vector autoregressions. Two of us have expressed elsewhere the view that nonstructural methods are not appropriate for policy evaluation because the errors and parameters in nonstructural models are complex (and unidentified) functions of the underlying structural errors and parameters (Cooley and LeRoy, 1983). See also Bennett McCallum (1982).

Our opinion is that these limitations on the scope of policy analysis do not represent genuine difficulties, but rather spurious puzzles. The problem goes back to Lucas's introduction of the concept of a policy "regime" as distinct from a policy variable, and to his representation of the former by a parameter. The key to getting our thinking unstuck is to respect the essential distinction between parameters as representing things which are assumed not to change, such as measures of preferences and technology, and variables as representing things which do, like policy regimes. Different policy regimes are then represented by different realizations of a random process (not by different values of a deterministic forcing variable, as in Keynesian policy evaluation, for then the Lucas critique would apply).

We have encountered the view that all this amounts to logic-chopping, and that no point of substance is involved. We disagree. Important macroeconomic questions have different answers depending on whether they are approached in the way we criticize or as we recommend. For example:

Should economists disqualify themselves from conducting policy evaluations of specific historical episodes? Contrary to Lucas, there is no reason in principle why economists should decline to analyze specific historical episodes—that is, should be unwilling to rank different policy sequences evolving out of a common past (of course, this is not to minimize the practical difficulties attending such an exercise).

Does rational expectations apply only in the "long run," or does it apply equally well in

[2] Lucas has invoked the distinction between risk and uncertainty, attributed to Frank Knight (1921), to express the view that agents cannot be represented as behaving rationally when confronted with discretionary policy changes. This attribution to Knight of the risk-uncertainty distinction as relating to whether or not the probabilistic calculus is applicable is incorrect (see our 1983 paper, Appendix). Knight was talking about market failure, a topic not related to the present discussion. The risk-uncertainty distinction as used by Lucas owes more to John Maynard Keynes (1921) than to Knight.

the "short run"? If policy is identified with a parameter which changes, then rational expectations will not apply following a policy change until learning processes have converged—that is, not until the indefinite future. Franco Modigliani (1977) argued from this that the rational expectations conclusions are irrelevant to the real world, and Sargent (1981) partly conceded the point by observing that rational expectations applies only after "agents have caught on to them" (p. 232). However, if policy is modeled as the realization of a sequence of random variables, there is in fact no reason to relegate the application of the rational expectations policy conclusions to the distant future.

Does the equilibrium-rational expectations perspective lead to a recommendation that government policy be bound to simple rules? The case for simple rules is sometimes based on the presumption that under "discretion" agents cannot be assumed to act rationally, or, specifically, to form rational expectations. Simple rules, then, are advocated on the grounds that this is the only policy environment we have any hope of analyzing.[3] Our opinion is that the conception that the rationality assumption is inapplicable in certain policy environments derives from the practice of representing policy regimes by fixed parameters, plus the observation that under frequent regime changes the fixity assumption is implausible. However, once we break ourselves of the habit of associating policy regimes with fixed parameters, it is seen that there is no justification for restricting the assumptions of rationality and rational expectations to certain policy environments and not others. Thus the argument for simple rules disappears.

A case for simple rules can only be based on a welfare analysis conducted using a model which is capable of representing individuals' (optimizing) behavior under either simple rules or complex response functions. If, as in our recommended modeling practice, parameters are reserved for measures of tastes and technology, and different policies are modeled as different realizations of a random process, then there is in principle no reason why such a welfare analysis cannot be undertaken.[4]

[3] Even if we accept the idea that agents' welfare is somehow connected with the simplicity of their decision problems, this does not create a presumption in favor of simple rules. The purpose of policy feedback rules (i.e., "discretion") is to offset exogenous shocks; to the extent that the government succeeds in doing so, it simplifies rather than complicates agents' decisions problems.

[4] These ideas are considered at greater length in our earlier paper.

REFERENCES

Cooley, Thomas F. and LeRoy, Stephen F., "Atheoretical Macroeconometrics," reproduced, University of California, Santa Barbara, 1983.

______, ______, and Raymon, Neil, "Modeling Policy Interventions," reproduced, University of California, Santa Barbara, 1983.

Fischer, Stanley, *Rational Expectations and Economic Policy*, Chicago: University of Chicago Press, 1980.

Keynes, John Maynard, *A Treatise on Probability*, New York: Harper and Row, 1921.

Knight, Frank C., *Risk, Uncertainty and Profit*, New York: Houghton Mifflin, 1921.

Lucas, Robert E., Jr., "Econometric Policy Evaluation: A Critique," in K. Brunner and A. Meltzer, eds. *The Phillips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Series on Public Policy, *Journal of Monetary Economics*, Suppl. 1976, 19–46.

______, "Rules, Discretion and the Role of the Economic Advisor," in S. Fischer, ed., *Rational Expectations and Economic Policy*, Chicago: University of Chicago Press, 1980.

______, *Studies in Business-Cycle Theory*, Cambridge: MIT Press, 1981.

______ and Sargent, Thomas J., *Rational Expectation and Econometric Practice*, Minneapolis: University of Minnesota Press, 1981.

McCallum, Bennett T., "Macroeconomics After a Decade of Rational Expectations: Some Critical Issues," Working Paper No. 1050, National Bureau of Economic Research,

1982.

Modigliani, Franco, "The Monetarist Controversy or, Should We Forsake Stabilization Policies?," *American Economic Review*, March 1977, *67*, 1–19.

Poole, William, "Macroeconomic Policy, 1971–75: An Appraisal," in S. Fischer, ed., *Rational Expectations and Economic Policy*, Chicago: University of Chicago Press, 1980.

Sargent, Thomas J., "Interpreting Economic Time Series," *Journal of Political Economy*, April 1981, *89*, 213–48.

Sims, Christopher A., "Policy Analysis with Econometric Models," *Brookings Papers on Economic Activity*, 1:1982, 107–64.

Taylor, John B., "Monetary Policy During a Transition to Rational Expectations," *Journal of Political Economy*, October 1975, *83*, 1009–21.

Controlling Contradictions Among Regulations

By LESTER B. LAVE*

Congress pursues externalities one at a time. The resulting legislation embodies the same sequential approach, instructing regulatory agencies to set and enforce standards for a single problem. Rarely, if ever, are agencies instructed or even permitted to account for contradictions with other federal legislation and rulemaking.

Much has been written about how an agency should behave in order to optimize social welfare in pursuing some particular objective (E. J. Mishan, 1976; Edith Stokey and Richard Zeckhauser, 1978). After accounting for uncertainty, the optimal scale of the program is the point where marginal benefit is just equal to marginal cost; the program should be done at this scale only if total benefits exceed total costs. The benefit-cost literature has debated the conditions required to maximize social welfare. The debate was recently rejuvenated by President Reagan's Executive Order 12291, requiring benefit-cost analysis for each major regulation; the alternative that maximizes net benefits must be chosen unless the statute specifies otherwise.

There is also a large literature on the political economy of the situation (James Wilson, 1980; S. G. Breyer, 1982). Neither Congress nor regulatory agencies act like philosopher kings, pursing social optima in a frictionless world of full information. Instead, decisions are sensitive to what data are available at the time action is taken, the income distribution implications of governmental actions, and institutional considerations including the current people in power.

A further potentially important difficulty that hasn't been treated in the literature is the possibility that various congressional goals and legislation may be directly or indirectly contradictory. Optimizing one externality assumes that it is independent of other externalities. This is unlikely. Insofar as the externalities are interdependent, the current system will produce suboptimization or outright contradictions. In what follows, I model the contradictions among regulations, show the errors, and then show how a wider perspective can help.

I. A Model of Contradictory Automobile Regulations

In 1966 Congress enacted the National Traffic and Motor Vehicle Safety Act to regulate the safety of automobiles. In 1970 Congress passed amendments to the Clean Air Act that required emissions of carbon monoxide and hydrocarbons be reduced by approximately 95 percent (90 percent for oxides of nitrogen), compared to an uncontrolled car. Finally, in 1975 Congress enacted the Energy Policy and Conservation Act requiring that the average new car sold in 1985 get 27.5 miles per gallon. Although the last act recognizes the possibilities of contradictions with previous acts, that is not dealt with in detail. In administering the act, the National Highway Traffic Safety Administration (NHTSA) has paid little attention to contradictions.

The automobile situation might be represented by the following:

$$(1) \qquad U(s, f, e, c, p),$$

where social utility (U) is a function of the attributes of each automobile: safety (s), fuel economy (f), emissions (e), comfort-performance (c), and price (p). All attributes are desirable except higher price.

To simplify, each of these attributes depends on three underlying attributes of the vehicle: weight (w), horsepower (h), and emissions control devices (d), each of which is the instrument for requiring safety, fuel economy, and emissions, respectively. In par-

[*] Professor of Economics and Public Policy, Carnegie-Mellon University, Graduate School of Industrial Administration, Schenley Park, Pittsburgh, PA 15213.

ticular, the relationships are shown as[1]

$$(2)\quad s(w);\ f(w,h,d);\ e(h,d);\ c(w,h,d);\ p(w,h,d).$$

Safety increases with weight, both because additional safety features add weight and because larger cars are somewhat safer than small ones. Fuel economy decreases with weight, horsepower, and emissions control. Emissions increase with horsepower and decrease with emissions control devices. Perception of comfort-performance is assumed to decrease with weight, other factors held constant. Comfort tends to increase with weight while performance increases with horsepower and decreases with emissions control. Finally, price increases with weight, horsepower, and emissions control.[2]

Thus, the social utility function can be rewritten as

$$(3)\quad U(s(w), f(w,h,d), e(h,d), c(w,h,d), p(w,h,d)).$$

The NHTSA is instructed by its legislation to increase highway safety. In this simplified model, increasing safety requires increasing vehicle weight (for example, by requiring more safety features). Were NHTSA to follow its legislation in an unthinking, inflexible manner, they would set the derivative to zero, as the following directs.

$$(4)\quad \operatorname*{Max}_{w} S:\ ds/dw = 0.$$

The NHTSA is not so foolish. Clearly, adding safety devices is subject to diminishing returns and, even for an agency pursuing only a single goal of safety, enough is enough. However, the judgement about how much safety is enough has changed from administration to administration. For example, President Ford ordered an elaborate social experiment to determine the effectiveness of air bags in practice. President Carter cancelled the experiment and ordered either air bags or passive seat belts on all cars. President Reagan ordered the requirement vacated (and the Supreme Court rescinded the revocation). Better decisions would result if Congress clarified NHTSA's goals. In particular, NHTSA could be instructed by Congress to examine the effects of its safety decisions on all automobile attributes, not just on safety. To do this, they would take the partial derivative of relation (3) with respect to weight and set this equal to zero, as shown in

$$(5)\quad \operatorname*{Max}_{w} U:\ \frac{\partial U}{\partial s}\frac{\partial s}{\partial w} + \frac{\partial U}{\partial f}\frac{\partial f}{\partial w} + \frac{\partial U}{\partial c}\frac{\partial c}{\partial w} + \frac{\partial U}{\partial p}\frac{\partial p}{\partial w} = 0.$$

Equations (4) and (5) represent a comparison between examining the effect of weight only on safety and examining the effect of weight on all vehicle attributes. Both partial derivatives in the first term are positive; the two partials in the next two terms are positive and negative, respectively. The two partials in the last term are negative and positive, respectively.[3] Thus, equation (5) will have the effect of tempering the optimal weight of a vehicle, compared to the solution to equation (4). This is not surprising, since weight is being modeled here as desirable only for safety.

The solution to equation (5) is quite different from that to equation (4). From a social viewpoint, equation (5) would be expected to provide a much higher level of utility, even though it represented a lower level of safety.

[1] The existence and nice properties of s, f, and e are shown in my 1981 paper.

[2] These statements assume other factors are held constant. Cars have become more fuel efficient and have lower emissions, compared with 1970. This was accomplished through innovation and higher manufacturing cost. Since 1974, manufacturers have made large investments in reducing vehicle weight, holding comfort-performance constant. Nonetheless, costs rise with increased weight, emissions, safety, and fuel economy held constant. The results would not change if an *R&D* or quality of manufacturing variable were added to each of the relationships in equation (2). Auto makers could be expected to invest in innovation and to increase the quality of manufacturing to the point where the marginal cost of more innovation and manufacturing quality would equal the savings in w, h, and d plus the increase in the desirability of the automobile.

[3] Weight tends to increase comfort and decrease performance.. Retaining simplicity leads to lumping these two together and assuming the partial derivative is negative.

However, in order to implement equation (5), NHTSA would have to know the numerical values of the other partial derivatives. Since these could be expected to vary with the weight, horsepower, and emissions control of vehicles, NHTSA would have to make some assumption about these. Alternatively, NHTSA could consult those who regulate emissions and fuel economy.

The Environmental Protection Agency (EPA) regulates emissions (actually, the agency does little more than implement the standards set by Congress). If EPA were trying to minimize emissions, it would increase emissions control devices. If EPA were instructed by Congress to account for the other consequences of its decisions, it would determine d by taking the partial derivative of relation (3) with respect to d and set this equal to zero, as shown in

$$(6)\qquad \underset{d}{\text{Max}}\, U: \frac{\partial U}{\partial f}\frac{\partial f}{\partial d} + \frac{\partial U}{\partial e}\frac{\partial e}{\partial d} + \frac{\partial U}{\partial c}\frac{\partial c}{\partial d} + \frac{\partial U}{\partial p}\frac{\partial p}{\partial d} = 0.$$

As with NHTSA, the EPA would have to know the values of the various partial derivatives, which depend on the levels of weight and horsepower set by other agencies.

Another part of NHTSA regulates fuel economy. If they were instructed only to maximize fuel economy, they would set horsepower to zero, an absurd result—that might account for congressional recognition in this act that there may be contradictions among regulations. If they were instructed to set horsepower by optimizing social utility, they would take the partial derivative of relation (3) with respect to horsepower, resulting in the following:[4]

$$(7)\qquad \underset{h}{\text{Max}}\, U: \frac{\partial U}{\partial f}\frac{\partial f}{\partial h} + \frac{\partial U}{\partial e}\frac{\partial e}{\partial h} + \frac{\partial U}{\partial c}\frac{\partial c}{\partial h} + \frac{\partial U}{\partial p}\frac{\partial p}{\partial h} = 0.$$

As before, they would have to know the values of weight and emissions control set by regulatory agencies.

Thus each agency would have to know the numerical values of partial derivatives whose values are controlled by federal regulatory agencies. My proposal is not to have the Office of Management and Budget draw up a list of assumptions that each agency should used. While this is a somewhat typical governmental solution, it is far from optimal. Instead, what is required is simultaneous optimization of relation (3) with respect to weight, horsepower, and emissions control, by simultaneously solving equations (5)–(7).

This formulation makes clear the limitations of the current statutory injunctions to each agency and of the strategy followed by Congress of attacking one problem at a time. At the very least, an agency ought to consider the implications of its regulatory actions on the other attributes of the product it regulates. More generally, all agencies that regulate a single product ought to formulate policy simultaneously, taking account of the values to be set by other regulatory agencies.

II. Implementing the Model for Automobiles

To show this optimization is possible and helpful, I will attempt to apply it crudely to the automobile. The problem is more general since there are myriad agencies regulating products, services, workplaces, etc. Whenever two agencies or two parts of one agency regulate the same area, simultaneous optimization is warranted.

Given the crudeness of this model, I will not attempt to solve the three equations simultaneously. Instead, I will try to implement equation (5). The marginal social utility of safety might be expressed in terms of the social value of preventing a premature death. A large literature attempts to clarify this notion and provide values. I will assume the value is $500,000 per premature death averted (Joanne Linnerooth, 1979; W. Kip Viscusi, 1983). Past data indicate that reducing vehicle weight from 4,500 to 3,000 pounds increases the chance of being killed or seriously injured in a crash from .04 to .06 (see my earlier paper). (I neglect the effect of in-

[4] Actually, NHTSA could set horsepower or weight to zero to maximize fuel economy. If they could regulate both instruments, they would consider equations (6) and (7) together.

creased vehicle weight on pedestrians and other vehicles.) Such a weight reduction for the entire fleet would be expected to result in 11,400 more fatalities each year, other factors held constant. Thus 7.6 fatalities or serious injuries would be expected for each 1 pound decrease in weight per year in all vehicles.

The social value of gasoline conservation is presumably more than the current cost of gasoline (T. J. Teisberg, 1981). Some approaches have put valuations to the marginal barrel of crude oil that are several times the current price. I assume the value of gasoline is $2.50 per gallon. A weight reduction from 4,500 pounds to 3,000 pounds would be expected to increase fuel efficiency by 5.7 miles per gallon, from 14.3 to 20 miles per gallon. For a fleet of 110 million cars, each going about 10,000 miles per year, the savings would be $36 million per 1 pound decrease in weight per year for all vehicles. (It is more difficult to provide estimates of the social value of comfort-performance or the extent to which this varies with weight. I will neglect this term.) An additional pound of weight probably costs $.25 or an annual cost per vehicle of $.06. Thus, for 110,000,000 vehicles roughly similar to the 1974 model, a weight increase of 1 pound on each vehicle would have instantaneous values shown in

$$(8)\quad \$500{,}000(7.6)-\$2.50(14{,}700{,}000) -\$.06(110{,}000{,}000)=\$3{,}800{,}000 -\$36{,}700{,}000-\$6{,}600{,}000=-\$39{,}500{,}000.$$

The immediate conclusion is that society would be better served by shaving weight off this vehicle than by adding additional safety features. In fact that has occurred, with a vast reduction in weight over time, and a consequent decrease in the inherent safety of the vehicle.

III. Implementation Problems

The above formulation assumes that various agencies will be attempting to maximize the same social welfare function. In fact, they are likely to give more weight to their own social problems; each agency could increase its institutional utility by presenting other agencies with misleading information about the functions, costs, or values of the variables to be evaluated. The literature on transfer pricing suggests some ways of attempting to elicit accurate responses.

More generally, current data are likely to provide an estimate of the incremental tradeoffs close to the current values, rather than some functional form likely to hold over a wide range. However, what is appealing about this formulation is that an agency could switch from pursuing a single objective in a narrow way (ds/dw) to a more general evaluation of the desirability of changing in the attribute under its control ($\partial s/\partial w$). If the values of the other partial derivatives are small, the more general approach can be dropped. In practice, no agency is likely to seek the trouble associated with simultaneous optimization. However, for particular situations, such as has occurred for the automobile, simultaneous optimization is the answer to the question of why each individually desirable regulation does not add up to a desirable outcome.

IV. Conclusion

The decade of the 1970's imposed environmental and health and safety regulation throughout the economy, adding to the economic regulation, equal opportunity regulation, and other prior government regulation. One part of EPA tells a company it cannot dump wastes into the river; another part forbids emissions into the air; a third part regulates dumping onto land. At the same time, the Occupational Safety and Health Administration requires that workers be protected against harmful exposures to the chemicals. Regulation of the automobile received national attention because of the depressed state of the industry. However, many other aspects of the economy are subject to contradictory regulations generated by the "one at a time" approach of Congress to externalities. Although the more general analysis proposed is more difficult to implement, the effort is necessary.

REFERENCES

Breyer, S. G., *Regulation and Its Reform*, Cambridge: Harvard University Press, 1982.

Lave, Lester B., "Conflicting Objectives in Regulating Automobiles," *Science*, May 22, 1981, *212*, 893–9.

Linnerooth, Joanne, "The Value of Human Life: A Review of the Models," *Economic Inquiry*, January 1979, *17*, 52–74.

Mishan, E. J., *Cost-Benefit Analysis: New and Expanded Edition*, New York: Praeger, 1976.

Stokey, Edith and Zeckhauser, Richard, *A Primer for Policy Analysis*, New York: W. W. Norton, 1978.

Teisberg, T. J., "A Dynamic Programming Model of the U.S. Strategic Petroleum Reserve," *Bell Journal of Economics*, Autumn 1981, *12*, 526–46.

Viscusi, W. Kip, *Risk by Choice*, Cambridge: Harvard University Press, 1983.

Wilson, James Q., *The Politics of Regulation*, New York: Basic Books, 1980.

Involuntary Unemployment as a Principal-Agent Equilibrium

By JAMES E. FOSTER AND HENRY Y. WAN, JR.*

Whether and why involuntary unemployment exists with persistence are issues of continuing debate. Advocates of the natural rate hypothesis deny that unemployment can be involuntary in models having rational agents, and favor structural reforms rather than discretional interventions in macroeconomic policies.[1] Keynesians would not so readily dismiss the presence of involuntary unemployment: instead they seek out microeconomic explanations and consider appropriate remedies for each separate source of the "malaise."[2]

Straightforward microeconomic explanations of involuntary unemployment, though, are not so easy to come by. As it is well known, the Walras equilibrium concept precludes persistent involuntary unemployment,[3] while search-theoretic explanations are usually identified with joblessness of a frictional and voluntary nature.[4] Explanations involving labor turnovers (for example, wages are kept above equilibrium to help recruitment and discourage quitting) might also be regarded as frictional and not involuntary.[5] Critics of the implicit contract model argue that the unemployment predicted is not only voluntary, but is also preventable by efficient contracts.[6]

Other theories tracing inflexible, high wages to their beneficial effects on the morale and efforts of workers are confronted with the possibility that firms might adopt performance contracts (i.e., incentive wage systems), including the piece rate, to sidestep this issue altogether.[7] The merits of the various arguments lie beyond our present scope.

The purpose of this paper is to observe that, in a principal-agent model with performance contracts, there may exist unemployment that is both persistent and involuntary in nature.

We consider a simple principal-agent model based on Leonid Hurwicz and Leonard Shapiro (1978) and on Milton Harris and Robert Townsend (1981). Firms control the means of production and hire workers to produce a single good. The output of a worker depends on the number of workers employed by the firm, the worker's effort level, and the worker's "status" (for example, his health, mood, etc.). The status of each worker is assumed to be identically and independently distributed. The firm can observe a worker's output, but not his effort, nor his status.

The firm chooses the number of workers hired and a contract for each worker, stipu-

*Krannert School of Management, Purdue University, West Lafayette, IN 47907, and Department of Economics, Cornell University, Uris Hall, Ithaca, NY 14853, respectively. We thank Jack Barron, James Moore, Paul Thomas, and Andrew Weiss for helpful comments. Any errors or shortcomings are ours alone.

[1] For example, see Milton Friedman (1968) and Robert Lucas (1977, 1978).

[2] See Robert Solow (1980) and the various contributions cited by him. On this view, if different causes of "wasteful" unemployment coexist, each should be separately analyzed and controlled; i.e., Occam's razor is inapplicable here. As an analogy, what were generically regarded as "fevers" in the last century are treated with different antidotes today.

[3] Lucas (1977), advocating an "equilibrium" theory of business cycles, would insist upon models that either are Walrasian or share many elements of the Walras framework. He cites Hayek: "By 'equilibrium theory' we here primarily understand the modern theory of general interdependence of all economic quantities, which has been most perfectly expressed by the Lausanne School of theoretical economics" (p. 7). Rationality of all individuals plays a central role in these models.

[4] We have in mind the important and numerous contributions evolved from the volume of Edmund Phelps et al. (1970).

[5] For example, see the penetrating analyses of George Akerlof (1976) and Andrew Weiss (1980).

[6] Unemployment in the implicit contract model of Costas Azariadis (1975), Martin Baily (1974), and Donald Gordon (1974) has been criticized by Robert Barro (1977) as preventable and by Oliver Hart (1983) as not being involuntary.

[7] Such explanations are discussed in Solow (1979), Roger Sparks (1982), Joseph Stiglitz (1982), and James Tobin (1972). Also see Wan (1973) for earlier literature and Takashi Negishi (1979) for an analytical review of the related issues. The criticism of such models was kindly provided to one of the authors by Walter Oi.

lating a (nonnegative) reward for each (nonnegative) output of the worker, possibly in some nonlinear manner.

Each worker has a utility function that depends positively on his reward and negatively on his efforts expended. He takes the number of coworkers and his contract as given, observes his status, and then chooses a utility-maximizing level of effort. So long as the worker's output is (continuously) increasing in effort, we can regard his "indirect" utility to be a function of his reward as well as his *output*, and parametrically dependent on the number of coworkers and his status.[8] With this in mind, we can regard the worker's choice variables to be his status-specific output levels instead of his effort levels.

The firm is assumed to know the worker's utility function and the underlying distribution of his status. The expected profit of a firm is the excess of outputs over rewards, averaged over all possible statuses and summed over all its workers. The problem facing a firm is to select the number of workers to hire and a contract for each worker, consistent with output responses maximizing its expected profit.

In Section I, profit-maximizing employment programs are explicitly derived for a parametric example with identical workers. Proposition 1 gives conditions under which there is an excess supply of labor when each firm adopts its profit-maximizing program. Moreover, the unemployment so obtained is "involuntary" in that each employed worker enjoys a strictly higher level of (expected) utility than every unemployed worker. The underlying reason for this is as follows. The profit-maximizing contract leaves a worker indifferent between working and shirking when in a less productive status, while inducing a strictly higher level of output (and utility) when the worker is in a more productive status. Thus the worker receives a strictly higher level of expected utility when employed. Yet the firm will not lower the wage schedule in response to an excess supply of labor. A wage cut at the level of output chosen in the less productive status would induce the worker to produce zero output; a wage cut at the more productive output level would lead the worker to produce the output of the less productive status. A firm's profit would fall if it were to lower the wage schedule in response to an excess supply of labor, even though each employed worker is strictly better off than his unemployed twin.[9]

Section II shows that the intuition behind this result continues to hold in the more general model: under quite minimal conditions, any unemployment that arises must be of an involuntary nature (see Proposition 2). A brief discussion of the main results is given in Section III.

I. Example

To make our exposition as transparent as possible, we shall consider a simple example with specific functional forms similar to Hurwicz-Shapiro and Harris-Townsend.

There are m identical firms producing the same (numeraire) commodity, and a set I of workers each having the same nonlabor income, preference, and ability at work. The number of workers is denoted by L.[10] The assumption of identical workers is not strictly necessary for our results, but as stressed by Takashi Negishi (p. 30) in a related context, it ensures that labor heterogeneity is not the cause of the involuntary unemployment.

The von Neumann-Morgenstern utility index of each individual i from I is of the form $u(r^i, z^i) = r^i - (z^i)^2$ where $r^i \geq 0$ is the reward to the individual, and $z^i \geq 0$ is his effort level. Each worker has two possible statuses of productivity, $k^i = 1$ or $k^i = 2$, with corresponding probabilities $p_1 > 0$ and $p_2 > 0$. The common production technology for each worker is represented by

$$(1) \qquad y^i = f(z^i; k^i, N) = a(k^i)e^{-bN}z^i,$$

where $a(k^i) > 0$ is the status-dependent pro-

[8] Compare with Hurwicz-Shapiro (p. 182).

[9] We are indebted to Andrew Weiss for this nice explanation.

[10] Strictly speaking, L is the measure of the set I. To avoid problems of indivisibilities, it is best to interpret I as the interval $[0, L]$.

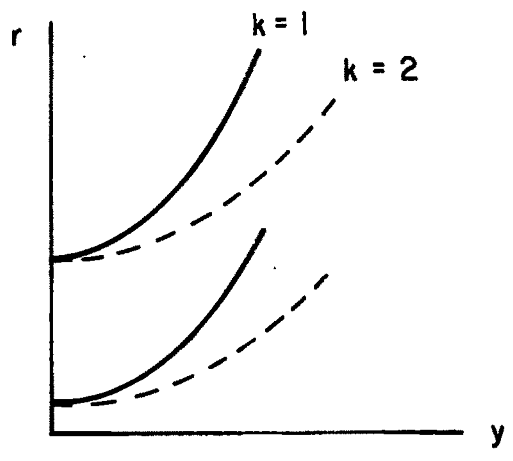


FIGURE 1

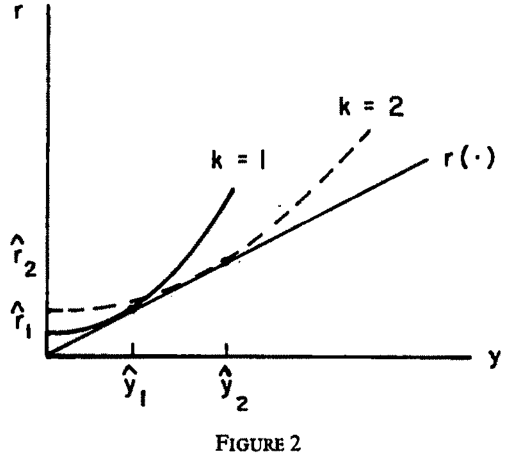


FIGURE 2

ductivity parameter, $N \geq 0$ is the number of workers hired by the firm, and $b > 0$ is a parameter representing scale diseconomies. Without loss of generality we take $a(1) < a(2)$; that is, status 2 is the more productive status.

Given k^i and N, we may invert f to obtain the effort-requirement function $g(y^i; k^i, N) = e^{bN}y^i/a(k^i)$, which indicates the amount of effort needed to produce a given output y^i. The "indirect" utility for individual i is then

$$(2)\quad v(y^i, r^i; k^i, N) = u(r^i, g(y^i; k^i, N))$$
$$= r^i - (e^{bN}y^i/a(k^i))^2.$$

Figure 1 depicts two sets of indifference curves for the indirect utility function, one set per status. The steeper slope of the status 1 indifference curves indicates that the worker must expend greater effort to produce the same output, hence must receive a higher reward to render him indifferent.

Being unable to observe a worker's status or monitor his effort level, the firm offers a performance contract $r^i(\cdot)$ which promises a nonnegative reward of $r^i(y^i)$ depending on the observed output y^i. A simple linear or "piece-rate wage" contract is depicted in Figure 2. The firm is assumed to choose an *employment program*, which specifies a level of employment N, and a performance contract $r^i(\cdot)$ for each worker i hired by the firm.[11]

Once an individual is hired by a firm as part of an employment program, his status-dependent utility may be expressed as a function of y^i alone, namely,

$$(3)\quad w(y^i; k^i, N) = v(y^i, r^i(y^i); k^i, N)$$
$$= r^i(y^i) - (e^{bN}y^i/a(k^i))^2.$$

The worker then chooses output level $\hat{y}_1^i$ for status 1 and $\hat{y}_2^i$ for status 2 satisfying

$$(4)\qquad w(\hat{y}_1^i; 1, N) \geq w(y^i; 1, N)$$

$$(5)\qquad w(\hat{y}_2^i; 2, N) \geq w(y^i; 2, N)$$

for all $y^i \geq 0$. Examples of such status-dependent utility maxima are given in Figure 2: output $\hat{y}_1^i$ leads to a reward of $\hat{r}_1^i$ where the contract touches the highest possible indifference curve in status 1, and similarly for $(\hat{y}_2^i, \hat{r}_2^i)$.

The status-dependent output responses of worker i lead to a level of expected profit (due to worker i) of

$$(6)\quad p_1(\hat{y}_1^i - r^i(\hat{y}_1^i)) + p_2(\hat{y}_2^i - r^i(\hat{y}_2^i))$$

[11] Implications of this assumption are discussed in Section III. Note that it is *not* assumed that all performance contracts offered by the firm are identical, although this may typically hold at equilibrium.

for the employer. The total expected profit associated with an employment program is found by summing (6) over all workers employed by the firm. An employment program which leads to the highest level of total expected profit is called a *profit-maximizing program* for the firm. We have the following result.

PROPOSITION 1: *If $m/2b < L$, then there will be unemployment when each firm chooses its profit-maximizing program. Further, the expected utility of each unemployed worker will be strictly less than the expected utility of every employed worker.*

To show this we will explicitly determine a profit-maximizing program for each firm. First, let us assume that an arbitrary employment level of $N > 0$ is given for a firm. Notice that the contracts or outputs of a worker's coworkers have no effect on the contract or output for that worker. Thus we can maximize total expected profit by maximizing the expected profit of each worker independently.

LEMMA 1: *Suppose N and the contracts $r^j(\cdot)$ of all workers $j \neq i$ at a firm are given. Then the contract $r_N(\cdot)$ defined by*

$$(7)\quad r_N(y) = \begin{cases} 0 & \text{for } 0 \le y < \hat{y}_1(N) \\ \hat{r}_1(N) & \text{for } \hat{y}_1(N) \le y < \hat{y}_2(N) \\ \hat{r}_2(N) & \text{for } \hat{y}_2(N) \le y; \end{cases}$$

$$(8)\quad \hat{y}_1(N) = \alpha_1 e^{-2bN}, \quad \hat{y}_2(N) = \alpha_2 e^{-2bN},$$
$$\hat{r}_1(N) = \beta_1 e^{-2bN}, \quad \hat{r}_2(N) = \beta_2 e^{-2bN};$$

(where $\alpha_1, \alpha_2, \beta_1, \beta_2 > 0$ are terms involving $a_1 = a(1)$, $a_2 = a(2)$, p_1, and p_2 as specified in the Appendix) leads to the highest expected profits from worker i. (For the proof of Lemma 1, see the Appendix.)

Since $r_N(\cdot)$ yields the highest profit from a given worker irrespective of the contracts of all other workers at the firm, it is clear that a program in which all contracts are $r_N(\cdot)$ must be profit maximizing for that N. By a simple calculation, the maximum total expected profit that can be achieved with N workers is $Ne^{-2bN}K$, where $K = p_1(\alpha_1 - \beta_1) + p_2(\alpha_2 - \beta_2) > 0$. Since this function achieves a unique maximum at $\hat{N} = 1/2b$, the employment program having $\hat{N}$ workers and identical contracts $\hat{r}(\cdot) = r_{\hat{N}}(\cdot)$ for all employees is a profit-maximizing program. Clearly, then, the total number of workers employed at all m identical firms is $m\hat{N} = m/2b$, when each adopts its profit-maximizing program. Hence, if $m/2b < L$ as assumed above, then $L - m/2b > 0$ of the workers must be unemployed.

Now an unemployed worker i receives the utility level associated with the pair $(r^i, z^i) = (0,0)$ in both statuses, namely $u(0,0) = 0$. An employed worker, by contrast, enjoys

$$(9)\quad v(\hat{y}_1, \hat{r}_1; 1, \hat{N}) = \hat{r}_1 - e\hat{y}_1^2/a_1^2$$
$$v(\hat{y}_2, \hat{r}_2; 2, \hat{N}) = \hat{r}_2 - e\hat{y}_2^2/a_2^2$$

in statuses 1 and 2, respectively, where $\hat{r}_k = \hat{r}_k(\hat{N})$ and $\hat{y}_k = \hat{y}_k(\hat{N})$ for $k = 1,2$. By evaluating (4) and (5) at $y^i = 0$ and $y^i = \hat{y}_1$, respectively, we obtain

$$(10)\quad \hat{r}_1 - e\hat{y}_1^2/a_1^2 \ge 0,$$
$$\hat{r}_2 - e\hat{y}_2^2/a_2^2 \ge \hat{r}_1 - e\hat{y}_1^2/a_2^2.$$

Since $a_1 < a_2$, it follows that $\hat{r}_2 - e\hat{y}_2^2/a_2^2 > 0$ which by $p_2 > 0$ implies that the expected utility of an employed worker is strictly greater than that of an unemployed worker. This establishes Proposition 1.

Figure 3 depicts the contract $\hat{r}(\cdot)$ from the above profit-maximizing program, and the status-dependent responses $\hat{y}_1$ and $\hat{y}_2$ of the worker. The contract offers a "base pay" of $\hat{r}(\hat{y}_1)$ when an output at least as great as $\hat{y}_1$ is achieved, with a "bonus" of $\hat{r}(\hat{y}_2) - \hat{r}(\hat{y}_1)$ for reaching $\hat{y}_2$. The class of profit-maximizing contracts consists of all contracts that (i) pass through $(\hat{y}_1, \hat{r}(\hat{y}_1))$ and $(\hat{y}_2, \hat{r}(\hat{y}_2))$, and ($ii$) lie nowhere above the status 1 and 2 indifference curves through $(\hat{y}_1, \hat{r}(\hat{y}_1))$. Note that there is a slight ambiguity in the definition of profit since the best responses of the workers are not unique; the worker may

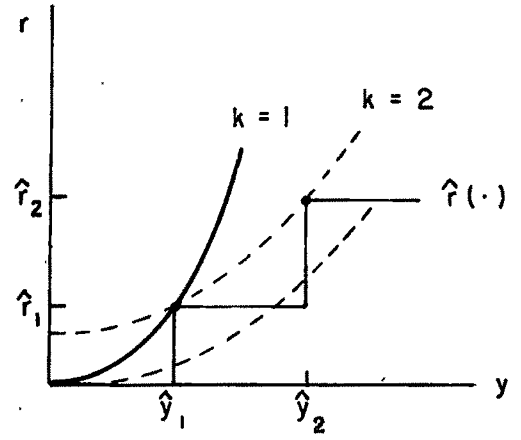


FIGURE 3

choose an output level of 0 in status 1, or $\hat{y}_1$ in status 2, without lowering his expected utility. To avoid this prospect, the firm can offer an "ε-optimal" contract in which a small amount of profit is forfeited to ensure that the unique response in status k is $\hat{y}_k$.

Remark: In the terminology of Harris-Townsend, the pairs $(\hat{y}_1, \hat{r}_1)$ and $(\hat{y}_2, \hat{r}_2)$ comprise a "parameter-contingent ($p.c.$) outcome" for the environment, satisfying the "self-selection" constraints

$$\text{(SS)}\quad v(\hat{y}_1, \hat{r}_1; 1, \hat{N}) \geq v(\hat{y}_2, \hat{r}_2; 1, \hat{N})$$
$$v(\hat{y}_2, \hat{r}_2; 2, \hat{N}) \geq v(\hat{y}_1, \hat{r}_1; 2, \hat{N})$$

and the "individual rationality" constraints

$$\text{(IR)}\quad v(\hat{y}_1, \hat{r}_1; 1, \hat{N}) \geq 0$$
$$v(\hat{y}_2, \hat{r}_2; 2, \hat{N}) \geq 0.$$

This can be seen in Figure 3 since $(\hat{y}_1, \hat{r}_1)$ lies on a status 1 indifference curve that is above $(\hat{y}_2, \hat{r}_2)$, and $(\hat{y}_2, \hat{r}_2)$ lies on the same status 2 indifference curve as $(\hat{y}_1, \hat{r}_1)$, verifying the (SS) constraints; while $(\hat{y}_1, \hat{r}_1)$ and $(\hat{y}_2, \hat{r}_2)$ lie on or above their respective state-dependent indifference curves passing through the origin, verifying the (IR) constraints.

II. A General Result

The above example illustrates a situation in which there exists persistent unemployment that is neither voluntary nor frictional. "Not frictional" is a by-product of the steady-state nature of the model. The employed and unemployed never change positions, and there are no workers caught between jobs. By "not voluntary," we mean that there are two identical individuals, one employed and the other jobless, where the former is better off than the latter in expected utility terms.

The two salient features of the example reinforce each other. On one hand, the persistence of the terms and levels of unemployment admits no tendency for change. On the other hand, identical persons enjoy different expected utility with "on-the-job" being favored over "on-the-dole."

Persistence of unemployment reflects adherence to profit-maximizing programs by firms. Barring wage concessions, firms would never hire more workers because of diminishing returns to employment. Further, any wage concession which might be offered by the jobless would always be resisted by firms. In a profit-maximizing program, a firm offers the lowest reward for each output response, until it becomes irreducible due to either the individual rationality or the self-selection constraints. Individuals never offer wage concessions which violate individual rationality: the utility of reward must at least cover the disutility of effort. Firms would never accept wage concessions that violate self-selection: such offers would be illusory as it would be incentive incompatible for workers to produce the particular outputs to which the concessions apply. Hence, any offered concession must be incredible while any credible concession would never be offered.

These conclusions are not limited to the simple two-status example presented above. Let us return to the general framework presented in the introduction, and assume that each identical worker has $s \geq 2$ possible statuses, where $p_k > 0$ for $k = 1, \dots, s$. Then employing the notation of Section I, the following proposition can be shown.

PROPOSITION 2: *Suppose that there exists a profit-maximizing program* $(\hat{r}(\cdot), \hat{N})$ *with status dependent responses* $\hat{y}_1, \dots, \hat{y}_s$. *If an output* $\hat{y}_k$, *can be produced in an alternative* *status $k'' \neq k'$ using less effort, then any unemployment must be involuntary.*

PROOF:

Denote the utility level of the unemployed by w_0. Since rewards are nonnegative, $w(0; k, \hat{N}) \geq w_0$, and since $\hat{y}_k$ is a utility-maximizing response in status k, we have $w(\hat{y}_k; k, \hat{N}) \geq w_0$, for each k.

Now suppose that $\hat{y}_{k'}$ can be produced in status $k'' \neq k'$ using less effort. Then since effort causes disutility,

$$(11) \quad w(\hat{y}_{k'}; k'', \hat{N}) > w(\hat{y}_{k'}; k', \hat{N}).$$

And since $\hat{y}_{k''}$ is a best response in status k'',

$$(12) \quad w(\hat{y}_{k''}; k'', \hat{N}) \geq w(\hat{y}_{k'}; k'', \hat{N}).$$

Thus, an employed worker obtains strictly higher utility than w_0 in status k'', and at least w_0 in every other status, so that the expected utility of the employed worker exceeds that of the unemployed.

Remark: Suppose that the worker is uniformly more productive in some status k'' than k', in that he can produce strictly more output at each positive effort level (see the above example). If the profit-maximizing program leads to a positive output in status k', then the hypothesis of Proposition 2 is satisfied.

The crucial aspect of our model is the presence of asymmetric information between worker and firm *favoring the worker*. This may be seen most clearly in terms of the individual rationality and self-selection constraints that must hold for a profit-maximizing contract. The individual rationality constraint implies that in any given status k, the worker's status-dependent response $\hat{y}_k$ is no worse than not working. By assumption, there are two statuses k' and k'' such that the output $\hat{y}_{k'}$ may be produced with less effort, and hence higher utility, in status k'' than in status k'. Since the firm is unable to observe the worker's status, it must abide by the self-selection constraint in status k'': the utility derived from producing $\hat{y}_{k''}$ must be at least as great as from producing $\hat{y}_{k'}$ in status k''. Thus, as compared to being unemployed, an employed worker receives strictly more utility in status k'', and no less utility in all other statuses, so that the expected utility of an employed worker is strictly higher than that of an unemployed worker.

III. Conclusion

We have shown that under certain circumstances it may be rational for a firm to pay its workers more than required to cover the disutility of labor effort, despite the presence of identical unemployed workers. That the accompanying unemployment is involuntary in nature is evident from the above discussion; yet we are not claiming to have obtained "Keynesian" unemployment. There remains a great deal of controversy surrounding the *General Theory*, and the chapter on involuntary unemployment is no exception. For instance, Robert Lucas (1978) dismisses Keynes' treatment of unemployment as "evasion and wordplay"; while Negishi infers from Chapter 2 a definition of involuntary unemployment similar to the one presented above. It is surely not our purpose to offer a novel interpretation of Keynes, nor to claim that the above model in any way captures the richness of the Keynesian "family of models" (Robert Clower and Axel Leijonhufvud, 1975, p. 182). Instead we have addressed the question of whether involuntary unemployment is a priori inconsistent with rational firms and workers. We have presented a preliminary answer in the form of a model with asymmetric information that admits involuntary unemployment in equilibrium.

Of course, our results may be sensitive to the precise form of the model we have adopted. For instance, we have assumed that a worker observes his status *after* being hired by the firm as part of an employment program. If the order of events were switched around so that the worker learned his status *before* being hired (or if renegotiation were allowed), intriguing questions would arise as to whether there are "rationing schemes" or other mechanisms which might better sort out high productivity workers from low, and lead to a higher expected profit for the firm. In the present paper, this is explicitly ruled out by an assumption which restricts the strategy of a firm to be an employment program.[12] It would be interesting to see how changing the model in this way, and allowing a richer strategy set for firms, might affect our results.

Another feature that should be noted is our requirement that contracts give a nonnegative reward for each output level. If firms are able to punish workers or force them to pay a penalty when output falls below a certain level, then one might find that the expected utility of those with jobs would be lowered to that of the jobless. Equivalently, involuntary unemployment in our sense may disappear if a firm can exact an "application fee" of such magnitude that any unemployed worker would be indifferent about whether he is hired or not.[13] Of course, this leads to serious questions about the enforceability of such labor contracts, both legally and practically, and in particular about what happens if a worker is forced below his budget constraint. In addition, one would expect the "reservation" level of utility (fixed at zero in our example) to be an endogenous variable of the model, determined by competition among firms and the relative scarcity of labor. Further work along these lines would surely be of interest.

Appendix

PROOF of Lemma 1:

Let N and the contracts of all other workers be given. Consider the contract $r_N(\cdot)$ defined in equations (7) and (8), where

$$\text{(A1)}\quad \alpha_1=\frac{p_1a_1^2a_2^2}{2\left(a_2^2-p_2a_1^2\right)},\qquad \alpha_2=\frac{a_2^2}{2};$$

$$\text{(A2)}\quad \beta_1=\left[\frac{p_1a_1a_2^2}{2\left(a_2^2-p_2a_1^2\right)}\right]^2,$$

$$\beta_2=\left(\frac{a_2}{2}\right)^2+\left(a_2^2-a_1^2\right)\left[\frac{p_1a_1a_2}{2\left(a_2^2-p_2a_1^2\right)}\right]^2.$$

For simplicity of notation, we denote $\bar{y}_k=\hat{y}_k(N)$ and $\bar{r}_k=\hat{r}_k(N)$ for $k=1,2$, and drop the superscript i in what follows. It is an easy matter to verify that $0<\alpha_1<\alpha_2$, and so $0<\bar{y}_1<\bar{y}_2$. Note that $\beta_1=(\alpha_1/a_1)^2$ and $\beta_2=(\alpha_2/a_2)^2+(a_2^2-a_1^2)(\alpha_1/(a_1a_2))^2$, so that $\bar{r}_1$ and $\bar{r}_2$ are strictly positive. The inequalities $\bar{r}_1<\bar{r}_2$ and $\beta_1<\beta_2$ will follow from Claim 3, below.

We shall show first that $\bar{y}_1$ is an optimal response in status 1, and $\bar{y}_2$ is an optimal response in status 2. Note that the contract $r_N(\cdot)$ specifies a constant reward on each of the intervals $[0,\bar{y}_1)$, $[\bar{y}_1,\bar{y}_2)$, and $[\bar{y}_2,\infty)$. Since effort generates disutility, the left endpoint in each interval is strictly preferred to the remaining points in the interval, irrespective of the status. Thus, we need only show $w(\bar{y}_1;1,N)\geq 0$ and $w(\bar{y}_1;1,N)\geq w(\bar{y}_2;1,N)$ to verify that $\bar{y}_1$ is an optimal response in status 1; and $w(\bar{y}_2;2,N)\geq 0$ and $w(\bar{y}_2;2,N)\geq w(\bar{y}_1;2,N)$ to show that $\bar{y}_2$ is an optimal response in status 2. This is done in the following four claims.

Claim 1: $w(\bar{y}_1;1,N)=0$.

$$w(\bar{y}_1;1,N)=\beta_1e^{-2bN}-\left(e^{bN}\alpha_1e^{-2bN}/a_1\right)^2$$
$$=e^{-2bN}\left(\alpha_1/a_1\right)^2-\left(e^{-bN}\alpha_1/a_1\right)^2.$$

[12] Under this assumption, Proposition 1 holds in the latter model as well, and has an interesting interpretation brought to our attention by Weiss. Since a worker observes his status before being hired by the firm, it is natural in this context to call the status of a worker his "type." Then there are two types of workers that the firm could hire, less productive and more productive, with the firm unable to identify a given worker's type. Proposition 1 shows that at the profit-maximizing employment program, the first type of worker is indifferent between working and being unemployed, while the second more productive type is strictly better off.

[13] For a discussion of application fees (or collateral requirements) in several related contexts, see Stiglitz and Weiss (1981; 1983, fn. 27).

Claim 2: $w(\bar{y}_1;1,N) > w(\bar{y}_2;1,N)$.

$$\begin{aligned}
w(\bar{y}_2;1,N) &= \beta_2 e^{-2bN} - \left(e^{bN}\alpha_2 e^{-2bN}/a_1\right)^2 \\
&= \left[\left(\frac{\alpha_2}{a_2}\right)^2 + \left(a_2^2 - a_1^2\right)\left(\frac{\alpha_1}{a_1 a_2}\right)^2 \right.\\
&\quad \left. - \left(\frac{\alpha_2}{a_1}\right)^2\right] e^{-2bN} \\
&= \left[\alpha_2^2 a_1^2 + \left(a_2^2 - a_1^2\right)\alpha_1^2 - \alpha_2^2 a_2^2\right] \\
&\quad \times e^{-2bN}/\left(a_1^2 a_2^2\right) \\
&= \left(a_2^2 - a_1^2\right)\left(\alpha_1^2 - \alpha_2^2\right) \\
&\quad \times e^{-2bN}/\left(a_1^2 a_2^2\right).
\end{aligned}$$

Since $a_2^2 > a_1^2$ and $\alpha_1^2 < \alpha_2^2$ we have $w(\bar{y}_2: 1, N) < 0$, which along with Claim 1 establishes the result.

Claim 3: $w(\bar{y}_2;2,N) = w(\bar{y}_1;2,N)$.

$$\begin{aligned}
&w(\bar{y}_2;2,N) \\
&\quad = \left[\left(\frac{\alpha_2}{a_2}\right)^2 + \left(a_2^2 - a_1^2\right)\left(\frac{\alpha_1}{a_1 a_2}\right)^2\right] e^{-2bN} \\
&\qquad - \left(e^{bN}\alpha_2 e^{-2bN}/a_2\right)^2 \\
&\quad = \left(a_2^2 - a_1^2\right)\left(\frac{\alpha_1}{a_1 a_2}\right)^2 e^{-2bN} \\
&\quad = \left(\frac{\alpha_1}{a_1}\right)^2 e^{-2bN} - \left(e^{bN}\alpha_1 e^{-2bN}/a_2\right)^2 \\
&\quad = w(\bar{y}_1;2,N).
\end{aligned}$$

Claim 4: $w(\bar{y}_2;2,N) > 0$.

$$\begin{aligned}
w(\bar{y}_2;2,N) &= w(\bar{y}_1;2,N) \\
&= \bar{r}_1 - \left(e^{bN}\bar{y}_1/a_2\right)^2 > \bar{r}_1 - \left(e^{bN}\bar{y}_1/a_1\right)^2 \\
&= w(\bar{y}_1;1,N) = 0,
\end{aligned}$$

since $\bar{y}_1 > 0$ and $a_2 > a_1$.

Now let $r(\cdot)$ be an alternative contract, with optimal responses y_1 and y_2. Denote $r_1 = r(y_1)$ and $r_2 = r(y_2)$. We shall show that the expected profit associated with $r(\cdot)$ can be no larger than the expected profits from $r_N(\cdot)$.

To be sure, the worker must find y_1 to be at least as preferred as 0 in state 1 under contract $r(\cdot)$, so that

$$\text{(A3)} \qquad r_1 \geq e^{2bN} y_1^2 / a_1^2.$$

Further, y_2 must be at least as preferred as y_1 in status 2 under $r(\cdot)$, so that

$$\text{(A4)} \quad r_2 - e^{2bN} y_2^2/a_2^2 \geq r_1 - e^{2bN} y_1^2/a_2^2,$$

and so

$$\text{(A5)} \quad r_2 \geq e^{2bN}\left(y_2^2/a_2^2 + y_1^2/a_1^2 - y_1^2/a_2^2\right).$$

In addition, recall that

$$\text{(A6)} \quad \bar{r}_1 = e^{2bN}\bar{y}_1^2/a_1^2,$$

$$\text{(A7)} \quad \bar{r}_2 = e^{2bN}\left(\bar{y}_2^2/a_2^2 + \bar{y}_1^2/a_1^2 - \bar{y}_1^2/a_2^2\right).$$

For $k=1,2$, denote $h_k = \bar{y}_k - y_k$ and $\mu_k = \bar{y}_k^2 - y_k^2$, and note that $\mu_k = 2h_k\bar{y}_k - h_k^2$. Then,

$$\begin{aligned}
\text{(A8)} \quad & p_1(\bar{y}_1 - \bar{r}_1) + p_2(\bar{y}_2 - \bar{r}_2) \\
&\qquad - p_1(y_1 - r_1) - p_2(y_2 - r_2) \\
&= p_1 h_1 + p_2 h_2 + p_1(r_1 - \bar{r}_1) + p_2(r_2 - \bar{r}_2) \\
&\geq p_1 h_1 + p_2 h_2 - p_1 e^{2bN}\mu_1/a_1^2 \\
&\quad - p_2 e^{2bN}\left(\mu_2/a_2^2 + \mu_1/a_1^2 - \mu_1/a_2^2\right) \\
&= p_1 h_1 - e^{2bN}\mu_1/a_1^2 + e^{2bN}\mu_1 p_2/a_2^2 \\
&\quad - p_2 e^{2bN}\mu_2/a_2^2 + p_2 h_2 \\
&= p_1 h_1 - e^{2bN}\mu_1\left(\frac{p_1}{2\alpha_1}\right) \\
&\quad - p_2 e^{2bN}\mu_2/a_2^2 + p_2 h_2 \\
&= p_1 h_1 - e^{2bN}h_1\bar{y}_1\left(\frac{p_1}{\alpha_1}\right) + e^{2bN}h_1^2\left(\frac{p_1}{2\alpha^1}\right) \\
&\quad + e^{2bN}\left(\frac{p_2}{2\alpha_2}\right)h_2^2 - p_2 h_2 \frac{e^{2bN}}{\alpha_2}\bar{y}_2 + p_2 h_2 \\
&= e^{2bN}\left(\frac{p_1}{2\alpha_1}h_1^2 + \frac{p_2}{2\alpha_2}h_2^2\right) \geq 0.
\end{aligned}$$

Thus, the expected profit associated with $r_N(\cdot)$ is highest among all contracts.

REFERENCES

Akerlof, George, "The Economics of Caste and of the Rat Race and Other Woeful Tales," *Quarterly Journal of Economics*, November 1976, *90*, 591–617.

Azariadis, Costas, "Implicit Contracts and Underemployment Equilibria," *Journal of Political Economy*, December 1975, *83*, 1183–202.

Baily, Martin N., "Wages and Employment Under Uncertain Demand," *Review of Economic Studies*, January 1974, *41*, 37–50.

Barro, Robert J., "Long-term Contracting, Sticky Prices and Monetary Policy," *Journal of Monetary Economics*, July 1977, *3*, 305–16.

Clower, Robert and Leijonhufvud, Axel, "The Coordination of Economic Activities: A Keynesian Perspective," *American Economic Review Proceedings*, May 1975, *65*, 182–88.

Friedman, Milton, "The Role of Monetary Policy," *American Economic Review*, March 1968, *58*, 1–17.

Gordon, Donald F., "A Neo-classical Theory of Keynesian Unemployment," *Economic Inquiry*, December 1974, *12*, 431–59.

Hart, Oliver D., "Optimal Labour Contracts Under Asymmetric Information: An Introduction," *Review of Economic Studies*, January 1983, *50*, 3–35.

Harris, Milton and Townsend, Robert M., "Resource Allocation Under Asymmetric Information," *Econometrica*, January 1981, *49*, 33–64.

Hurwicz, Leonid and Shapiro, Leonard, "Incentive Structures Maximizing Residual Gain Under Incomplete Information," *Bell Journal of Economics*, Spring 1978, *9*, 180–91.

Keynes, John M., *The General Theory of Employment Interest and Money*, New York 1936.

Lucas, Robert E., Jr., "Understanding Business Cycles," in Karl Brunner and Allan H. Meltzer, eds., *Stabilization of the Domestic and International Economy*, Vol. 5, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl. 1977, 7–29.

______, "Unemployment Policy," *American Economic Review Proceedings*, May 1978, *68*, 353–57.

Negishi, Takashi, *Microeconomic Foundations of Keynesian Macroeconomics*, Amsterdam: North-Holland, 1979.

Phelps, Edmund S. et al., *Microeconomic Foundations of Employment and Inflation Theory*, New York: Norton, 1970.

Solow, Robert M., "Another Possible Source of Wage Stickiness," *Journal of Macroeconomics*, Winter 1979, *1*, 79–82.

______, "On Theories of Unemployment," *American Economic Review*, March 1980, *70*, 1–10.

Sparks, Roger W., "A Model of Unemployment and Wage Rigidity," unpublished paper, University of California-Davis, 1982.

Stiglitz, Joseph E., "The Wage-Productivity Hypothesis: Its Economic Consequences and Policy Implications," paper presented at the annual meeting of the American Economic Association, 1982.

______ **and Weiss, Andrew,** "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, June 1981, *71*, 393–410.

______ **and** ______, "Incentive Effects of Terminations: Applications to the Credit and Labor Markets," *American Economic Review*, December 1983, *73*, 912–27.

Tobin, James, "Inflation and Unemployment," *American Economic Review*, March 1972, *62*, 1–18.

Wan, Henry Y., Jr., "A General Theory of Wages, Employment, and Human Capital—An Application of Semi-Competitive Equilibrium," Department of Economics Working Paper No. 51, Cornell University, 1973.

Weiss, Andrew, "Job Queues and Layoffs in Labor Markets with Flexible Wages," *Journal of Political Economy*, June 1980, *88*, 526–38.

The Theory of Trade in Middle Products: An Extension

By JEFFREY A. FRANKEL*

In a recent thought-provoking contribution to this *Review*, Kalyan Sanyal and Ronald Jones suggest a view of international trade in which all trade, "even for commodities which physically appear in final form in the world market," is thought of as consisting of intermediate inputs or "middle products" (1982, p. 16). After the commodity is shipped, the importer must put in his value-added, even if it consists of no more than local transportation and retailing services, before the good can be sold to consumers. This note points out an interesting implication of that view of trade for standard sticky-price macroeconomic models.[1] The implication is that direct short-run effects of exchange rate changes on the aggregate price level, and therefore on the demand for money and the supply of labor, are ruled out. This in turn rules out the often-heard claim that when monetary contraction in, say, the United States causes the dollar to appreciate against European currencies, it will be transmitted, via an increase in the European price level, as a contraction in Europe.

I begin by reviewing a conventional view of pricing in world goods markets: for a large class of commodities, prices are invoiced—and in the short run are fixed—in the currency of the country selling them. The reasons for the stickiness of these goods prices are not completely understood but obvious possible factors include imperfect information, costs to changing prices, inertia, explicit contracts and implicit contracts. These commodities are called "customer" goods by Arthur Okun (1975), "fixprice" goods by John Hicks (1974), and "Tradables I" by Ronald McKinnon (1979). Their existence is supported by the empirical failure of the Law of One Price for two countries' closely matched categories of manufactured goods, in studies by Peter Isard (1977) and Irving Kravis and Robert Lipsey (1977). The second class of traded commodities is called "auction" goods by Okun, "flex-price" goods by Hicks, and "Tradables II" by McKinnon. These goods, consisting of homogeneous basic commodities like oil, have prices that are determined by supply and demand on world markets. Perfect substitutability enforces the Law of One Price internationally for them. They are largely used as intermediate inputs in the manufacturing process, though they may also enter final consumption directly, especially in the case of food.

This conventional view implies that a country's Consumer Price Index is the weighted average of three kinds of goods: domestically produced fixed-price goods (including nontraded goods, if any), with prices that are sticky in domestic currency; foreign-produced fixed-price goods, with prices that in domestic currency vary one-for-one with the exchange rate; and flex-price goods, with prices that in domestic currency also vary with the exchange rate (one-for-one *if* the domestic country is small in this market, otherwise somewhat less). An increase in the exchange rate will instantaneously raise the *CPI* by an amount depending on the shares of consumption occupied by the last two kinds of goods.

Now I come to the point of this paper. If the Sanyal and Jones view is correct, *all goods at the stage when they are sold to consumers belong in the category of domestically produced fix-price goods*. Consumers do not line up on the docks to buy imports, whether foreign-produced fix-price goods or flex-price goods. Rather, these imported goods are bought by domestic firms who put in some value-added before selling them to domestic consumers at prices that are sticky. It follows that, whatever it is that causes the

*Department of Economics, University of California, Berkeley, CA 94720.

[1] Other implications of the Sanyal-Jones view in specific macro models are developed by Neil Bruce and Douglas Purvis (forthcoming), and by Jones and Purvis (1981).

prices of domestically produced goods to be sticky, *all* components of the *CPI* fall in this category. The exchange rate is not reflected at all, in the short run, in the domestic price level.

Why is the question important, whether the exchange rate affects the *CPI* in the very short run? One immediate application is the determination of inflation. If some components of the domestic price level respond instantaneously to the exchange rate, a domestic monetary contraction that causes the currency to appreciate, for example, will bring the price level down more quickly under floating exchange rates than under fixed exchange rates.[2]

Another important application concerns effects on output in open economies. Most commonly, one expects a depreciation of the domestic currency to *add* to the demand for domestic goods through the effect of improved competitiveness on the trade balance. But as pointed out by Richard Cooper (1971), and Krugman and Lance Taylor (1978), a depreciation can have a *contractionary* effect through a number of channels. Two of them follow from the positive effect on the domestic price level. First, the increase in the price level will raise nominal money demand, or reduce real money supply, especially if the price index that is considered relevant for the money market is the *CPI*.[3] This will exert a contractionary effect on the demand for domestic goods. Second, to the extent that domestic wages are rigid in real terms, for example, due to indexed wage contracts, the increase in the *CPI* will raise nominal wages and therefore raise real wages in terms of the domestically produced fix-price goods. This will exert a contractionary effect on the demand for labor and the corresponding supply of output in this sector.

Thus a foreign monetary contraction, which causes the foreign currency to appreciate against the domestic one, can be transmitted to the domestic country as a contraction. It also follows that a foreign fiscal contraction, which lowers foreign interest rates and so causes the foreign currency to depreciate, will be transmitted to the domestic country as an expansion because the domestic *CPI* will fall. These two channels, the effects of the exchange rate on money demand and wages, reverse Robert Mundell's famous results (1964) on transmission in a two-country model with capital mobility. They appear in papers by, among others, William Branson and Willem Buiter (1983), F. R. Casas (1975), Jeffrey Sachs (1980), Buiter and Marcus Miller (1982), and Rudiger Dornbusch (1983). But the two channels are open to question, at least as propositions about the very short run, if we adopt the macroeconomic corollary of the Sanyal and Jones view. If the exchange rate has no instantaneous effect on the price level, it can have no instantaneous effect on nominal money demand or nominal wages.[4]

How plausible is the argument that exchange rate changes do not affect the domestic-currency prices of imported goods in the short run? Surely an appreciation of the dollar against the yen is reflected in the U.S. currency price of a Mazda even if it is not reflected in the domestic currency price of a Ford? Casual empiricism suggests that the answer may be "no." The U.S. dealers who

[2] This point was made by Rudiger Dornbusch and Paul Krugman (1976). Willem Buiter and Marcus Miller (1982) argue that this effect holds only in the short run if sticky prices adjust over time. But the present paper is only concerned with the short run.

[3] It is an open question which price index is the proper one. Stephen Goldfeld (1973), for example, favors the *GNP* deflator. If the demand for money arises from transactions models, one can make an argument for using the Producer Price Index. While consumers hold money to transact in only final goods, firms hold money to transact in intermediate inputs as well, and the total number of transactions undertaken by firms is greater. But even if we were to use the *PPI* or *GNP* deflator, the exchange rate would still have an effect to the extent that flex-price goods, the prices of which are determined on world markets, are important.

[4] The question of how a U.S. monetary contraction is transmitted to Europe was of obvious practical relevance in the years 1980–82. How do we explain the large recession in Europe that paralleled that in the United States, if the effects of the dollar appreciation on European money demand and wages are ruled out? (Note that one of the most important channels that I have left out, the Laursen-Metzler effect, goes the other way: the adverse shift in the Europeans' terms of trade causes them to reduce savings and raise expenditure, implying an expansion in Europe.) Probably the best answer is that the Europeans contracted at the same time as the United States did, in order to prevent their currencies from depreciating any more than they did, and to fight inflation.

sell Japanese cars have list prices fixed in dollars. The U.S. consumer who watches the dollar soar against the yen and rushes down to his local Mazda dealer for a bargain, is likely to be disappointed. An exchange rate fluctuation opens up a gap between the price in Japan and the price in the United States. There is an obvious incentive to arbitrage. And the company will eventually move to close the gap, especially if it sees demand rising in the low-priced country and falling in the high-priced country. But there is no evidence that arbitrage is in practice able to close the gap in the short run.

It would be desirable to compare empirically the prices of imported goods with the prices of the physically identical goods in the country of production, and to see if the exchange rate can explain the gap. In the meantime, it appears at least possible that the absence of any short-run effect of the exchange rate on the domestic price level is more than just an interesting theoretical implication of the Sanyal and Jones paper; it may be an accurate description of actual behavior as well.

REFERENCES

Branson, William and Buiter, Willem, "Monetary and Fiscal Policy with Flexible Exchange Rates," in J. Bhandari and B. Putnam, eds., *Economic Interdependence and Flexible Exchange Rates*, Cambridge: MIT Press, 1983.

Bruce, Neil and Purvis, Douglas, "The Specification and Influence of Goods and Factor Markets in Open Economy Macroeconomic Models," in P. Kenen and R. Jones, eds., *North-Holland Handbook of International Economics*, forthcoming 1984.

Buiter, Willem and Miller, Marcus, "Real Exchange Rate Overshooting and the Output Cost of Bringing Down Inflation," *European Economic Review*, May/June 1982, *18*, 85–123.

Casas, F. R., "Efficient Macroeconomic Stabilization Policies Under Floating Exchange Rates," *International Economic Review*, October 1975, *16*, 682–98.

Cooper, Richard, "Currency Devaluation in Developing Countries," *Princeton Essays in International Finance*, No. 86, June 1971; Reprinted in J. Letiche, ed., *International Economic Policies and Their Theoretical Foundations*, New York: Academic Press, 1982.

Dornbusch, Rudiger, "Flexible Exchange Rates and Interdependence," *IMF Staff Papers*, March 1983, *30*, 3–30.

______ **and Krugman, Paul,** "Flexible Exchange Rates in the Short Run," *Brookings Papers on Economic Activity*, 3:1976, 537–84.

Goldfeld, Stephen, "The Demand for Money Revisited," *Brookings Papers on Economic Activity*, 3:1973, 577–638.

Hicks, John R., *The Crisis in Keynesian Economics*, New York: Basic Books, 1974.

Isard, Peter, "How Far Can We Push the 'Law of One Price'?," *American Economic Review*, December 1977, *67*, 942–48.

Jones, Ronald and Purvis, Douglas, "International Differences in Response to Common External Shocks: The Role of Purchasing Power Parity," paper presented at the Fifth International Conference of the University of Paris-Dauphine, May 1981.

Kravis, Irving and Lipsey, Robert, "Export Prices and the Transmission of Inflation," *American Economic Review Proceedings*, February 1977, *67*, 155–63.

Krugman, Paul and Taylor, Lance, "Contractionary Effects of Devaluation," *Journal of International Economics*, August 1978, *8*, 445–56.

McKinnon, Ronald, *Money in International Exchange*, New York: Oxford University Press, 1979, ch. 4.

Mundell, Robert, "A Reply: Capital Mobility and Size," *Canadian Journal of Economics and Political Science*, August 1964; adapted in R. Mundell, *International Economics*, New York: Macmillan, 1968, Appendix to ch. 18.

Okun, Arthur, "Inflation: Its Mechanics and Welfare Costs," *Brookings Papers on Economic Activity*, 2:1975, 351–90.

Sachs, Jeffrey, "Wages, Flexible Exchange Rates and Macroeconomic Policy," *Quarterly Journal of Economics*, June 1980, 731–47.

Sanyal, Kalyan and Jones, Ronald, "The Theory of Trade in Middle Products," *American Economic Review*, March 1982, *72*, 16–31.

Credible Commitments: Further Remarks

By Oliver E. Williamson*

I distinguish between unilateral and bilateral trading relations in an earlier article in this *Review* (1983) in which I examined the use of hostages (or their commercial equivalents) to support exchange. The purposes of this note are 1) to confirm the optimality of reciprocal trading by explicitly displaying the combined net benefit relations, and 2) to acknowledge a previously unremarked complication that arises when nonsalvageable assets are used as hostages in a unilateral trade.

I. Introduction

As in the earlier paper, risk neutrality is assumed and suppliers are assumed to be willing to produce to any contract for which expected break even can be projected. As before, orders are placed in the first period and all nonsalvageable costs needed to support production are incurred at that time. The same stochastic demand for product is assumed to obtain in the buyer's market; namely, the gross value realized by a buyer upon resale in the final product market is uniformly distributed over the interval 0 to 1, an identical quantity being demanded at every price, which quantity is assumed for convenience to be unity. The buyer decides to confirm or cancel the order upon being apprised of demand realizations in period 2.

Two technologies are considered and several different contracts are considered. The technologies and contracts of special interest to this paper are the following:

T_1: This is the general purpose technology, all advance commitments of which are salvageable, the unit operating costs of which are v_1;

T_2: This is the special purpose technology, the nonsalvageable value of advance commitments of which are k and the redeployable unit operating costs of which are v_2;

CII: The producer makes the specific asset investment himself and receives a payment of $\bar{p}$ in the second period if the buyer confirms the order but nothing otherwise;

CIII: The producer makes the specific asset investment himself and receives $\hat{p}$ from the buyer if the buyer confirms the order, is paid αh, $0 \leq \alpha \leq 1$, if the order is cancelled while the buyer pays $\hat{p}$ upon taking delivery and experiences a reduction in wealth of h if second period delivery is cancelled.

Delivery will be accepted whenever the net benefits realized by the buyer are nonnegative. Letting p be demand realization in the second period, the net benefits of taking delivery under contract II will be p less the price $\bar{p}$ that is paid to the supplier. The corresponding net benefits of contract III are p less $\hat{p}$ plus the avoided costs of cancellation, h. Accordingly, whereas $b_{\text{II}} = p - \bar{p}$, $b_{\text{III}} = p - \hat{p} + h$. If $\hat{p} = v_2 + k$ and $h = k$, this last becomes $b_{\text{III}} = p - v_2$, whence product will be accepted under contract III so long as realized demand price exceeds marginal cost.

II. Reciprocal Trade

Firms that are engaged in tied bilateral trading are confronted with a somewhat more complicated net benefit calculation. In deciding whether to take delivery or cancel an order, a firm needs to consider not merely the net gain from procurement but must also consider the net gain from supply. Let the net gain from buying and selling product be given by b_B and b_S, respectively. The combined gain from observing reciprocity is then given by $b_R = b_B + b_S$. Let $\hat{p}$ be the price at which product is traded. Then the net benefits upon taking delivery in the purchase market will be given by

$$(1) \qquad b_B = p - \hat{p},$$

*Gordon B. Tweedy Professor of the Economics of Law and Organization, Yale University, New Haven, CT 06520.

while net benefits from the simultaneous sale of product (given that specific assets in amount k have already been sunk) are given by

$$(2) \qquad b_S = \hat{p} - v_2.$$

The net benefits of noncancellation—that is, of continuing reciprocal trade (given that one's trading counterpart does not renege)—are then

$$(3) \quad b_R = (p - \hat{p}) + (\hat{p} - v_2) = p - v_2,$$

which will be positive so long as demand realization in the purchase market exceeds the marginal cost of own-production.

Although the specific asset term, k, appears nowhere in these expressions, this does not imply that it is irrelevant. For one thing, the expected net benefits of reciprocity will be positive only if the probability of trade under the reciprocal trading criterion (namely, $1 - v_2$) times the expected gain from remunerative exchange $(1 - v_2)/2$ exceeds the value of nonsalvageable assets, k. Thus the inequality $(1 - v_2)^2/2 - k > 0$ must be satisfied. (In the limit, as v_2 approaches 0, a net gain from reciprocity will obtain only if $k \leq 1/2$.)

More significant is the fact that only if specific assets are committed in support of the exchange will the benefits from the sale of product be given by $b_S = \hat{p} - v_2$. If, for example, one of the parties to the exchange were to employ the general purpose technology T_1 instead, the net benefits from supplying product for which $\hat{p}$ is received would be $b'_S = \hat{p} - v_1$. The criterion for assessing whether to cancel or not would then be $b'_R = p - v_1$, which would call for cancellation in demand states where $p < v_1$. One party to the bilateral exchange would thus find cancellation attractive under circumstances where the other, because it has made specific asset investments, would want product to be traded. The symmetrical exposure of specific assets avoids this result.[1]

III. Unilateral Trade

Reciprocal trade has the interesting property that the hostages created to realign incentives are never exchanged. The expropriation hazards to which hostage creation is otherwise subject (see my earlier article, pp. 526–27) are thereby vitiated. Unilateral trade can likewise benefit from hostage creation. But the issues here are somewhat more complicated than my original discussion indicates. The buyer's incentives will only be partially realigned if transaction specific investments (in amount $k''' = k$) in successor stage activity are incurred if these assets are not turned over to the supplier upon order cancellation.[2]

To be sure, suppliers who make specialized investments at a buyer's behest will always sell products on more favorable terms to those buyers who make nonsalvageable investments in successor stage activities, *ceteris paribus*. Since buyers who make such investments will thereafter confirm orders in more adverse-demand states than those who do not, such investments constitute credible commitments. But the added credibility thereby realized will rarely reduce order cancellation hazards to zero. The price at which product is supplied will necessarily be adjusted to reflect whatever degree of uncompensated risk remains. As a consequence, the full optimality reached by reciprocal trading will not be realized by unilateral trading if hostages are not transferred upon cancellation.

This is not, however, to say that hostage transfers should always attend unilateral trading. Such a rule suffers from the aforementioned expropriation hazard. More generally, all feasible trading alternatives may be flawed. A comparative institutional assessment of the main organizational alterna-

[1] Symmetry is a sufficient but not a necessary condition for the parties to assess the net benefits of noncancellation identically. Other trading relations with this same property might conceivably be crafted. If, for example, one of the parties invests more in specific assets than does the other, then an identical assessment of the net benefits of noncancellation could obtain if the party with the lesser degree of asset investment were to sell in a final product market where the distribution of demand, to be denoted by p_1, is such that $p_1 - v_1 = p - v_2$.

[2] To serve as compensation, of course, these assets must be fully valued ($\alpha = 1$) by the supplier.

tives would presumably include consideration of the following:

1) Full compensation for order cancellation, in which event buyers are exposed to expropriation hazards;

2) Buyers invest in specific assets but refuse compensation, which creates credible commitments but exposes suppliers to the risk of uncompensated losses;

3) A compromise whereby suppliers create credible commitments and make partial but incomplete hostage payments upon order cancellation;

4) Expand the contractual relation by developing suitable reciprocity arrangements; and

5) Consolidate the transaction under common ownership, which is the vertical integration alternative.

Whether options 4 or 5 are feasible will vary with the circumstances. Here, as elsewhere, informed choice requires the type of detailed knowledge of the institutional realities of economic life to which Tjalling Koopmans referred (1957, p. 145). The attributes of the trading parties, the technologies to which they have access, and the markets in which they operate all need to be assessed.

REFERENCES

Koopmans, Tjalling C., *Three Essays on the State of Economic Science*, New York: McGraw-Hill, 1957.

Williamson, Oliver E., "Credible Commitments: Using Hostages to Support Exchange," *American Economic Review*, September 1983, *73*, 519–40.

How General is the Case for Unilateral Tariff Reduction?

By Paul Wonnacott and Ronald Wonnacott*

In this *Review* (1981), we attacked a proposition (P1) of Eitan Berglas (1979) and others, that unilateral tariff reduction (*UTR*) is *necessarily* superior to customs union (*CU*), provided scale economies and changes in terms of trade are ruled out. We put forward a much weaker proposition (P2) (p. 706) that *UTR* is sometimes superior, sometimes inferior. In his 1983 paper, (p. 1142), Berglas concedes that our 1981 Figure 2 (which he calls example E1) illustrates *CU* superiority. He thereby concedes our main point: P2 is correct; P1 is not.

A question remains: how interesting is the domain where *UTR* is superior?[1] We argued (1981) *UTR* is superior under narrow assumption A1 (partner *B*'s tariffs can be ignored) or A2 (no tariff by outsider *C* nor transport costs in trade with *C*). But a *CU* is superior in our main example (Figure 2), where neither A1 nor A2 holds, and partners *A* and *B* trade with mutual benefit in the price wedge between *C*'s import and export prices.

In addition to wrongly claiming we were illogical and incorrect,[2] Berglas argues (1983) that there are two other assumptions in his 1979 paper which, taken together, are also sufficient to establish *UTR* superiority. However, he misstates them. They are not, as he says, A3 (the *CU* does not affect the direction of trade) plus A4 (all three countries trade). Instead, they are A3 plus a much more restrictive A5: *C* trades *every* good with the *CU*. (This assumption, in Berglas, 1979, Figure 1 and Table 1, is distinguishable from A4 only when there are more than two goods.) Why must A5 be assumed, not just A4? Without A5, trade between *A* and *B* can occur in some goods within *C*'s price wedge. In short, Berglas establishes *UTR* superiority by A5, assuming that *A* and *B* can't trade in the wedge where *CU* provides mutual benefits. (Our wedge becomes increasingly important in the *n*-good case. Consider cement, for example.)

Moreover, Berglas's A3 rules out changes in trade patterns, and therefore Viner's concepts of trade diversion and trade creation which introduced the modern *CU* debate. Like a case based on A1 or A2, the A3 + A5 case for *UTR* superiority is not interesting. Even if it were, one more special case does not establish the general principle that *UTR* is necessarily superior to a *CU*, any more than one more example where protection raises welfare would establish a general proposition that protection necessarily raises welfare.

*Departments of Economics, University of Maryland, College Park, MD 20742, and University of Western Ontario, London ON N6A 3K7 Canada, respectively.

[1] By *UTR* superiority, we mean *UTR* is equal to or better than *CU*.

[2] First, he says we made a "logical mistake" in not proving "that A1 and A2 were necessary" for *UTR* superiority (1983, p. 1141). We didn't prove it because our argument didn't require it. Our argument hinged on A1 *or* A2 being *sufficient*. (See our earlier Figure 1, p. 707, where A2 is sufficient in the absence of A1.) Second, he states that we are "incorrect" in asserting that A1 and A2 are in the literature. But we showed (1981, p. 705) that A1 *was* made by other authors. And A2 *was* in Berglas (1979), as he acknowledges (1983, p. 1141, fn. 1).

REFERENCES

Berglas, Eitan, "Preferential Trading Theory: The *n* Commodity Case," *Journal of Political Economy*, April 1979, *87*, 315–31.

______, "The Case for Unilateral Tariff Reductions: Foreign Tariffs Rediscovered," *American Economic Review*, December 1983, *73*, 1141–42.

Wonnacott, Paul and Wonnacott, Ronald, "Is Unilateral Tariff Reduction Preferable to a Customs Union? The Curious Case of the Missing Foreign Tariffs," *American Economic Review*, September 1981, *71*, 704–14.

The Misuse of Accounting Rates of Return: Comment

By IRA HOROWITZ*

Consider the plight of I. Napfkavitch, the plaintiff's economic expert in an antitrust suit, who having concluded his direct testimony is subjected to the following cross examination.

Attorney: "Dr. Napfkavitch, have you ever heard of Professor Franklin M. Fisher?"

Napfkavitch: "Yes."

Attorney: "And is Professor Fisher a highly respected economist?"

Napfkavitch: "Yes."

Attorney: "Let me read something to you, Dr. Napfkavitch: '...examination of absolute or relative accounting rates of return to draw conclusions about monopoly profits is a totally misleading enterprise.' Do you agree or disagree with that statement, Dr. Napfkavitch?"

Having based a portion of his testimony on the fact that the defendant corporation, which historically has held a 40 percent market share in a stipulated relevant market, has commonly earned in the neighborhood of 30 percent after-tax profit on average stockholder's equity (in contrast to seemingly comparable corporations housed in the defendant's four-digit SIC industry who have commonly earned 10 to 15 percent), the esteemed Napfkavitch is forced to concede that on the whole he disagrees with the statement.

Attorney: "And, Dr. Napfkavitch, do you know who made that statement with which you so cavalierly disagree?"

Napfkavitch: "I don't recall for sure, but I do seem to recall reading it and I think I can guess. In any event, I'm positive you're about to tell me."

Attorney: "Professor Franklin M. Fisher, whom you've acknowledged to be a highly respected economist, made that statement, Dr. Napfkavitch, in a recently published paper that he coauthored with John J. McGowan. That paper appeared in the *American Economic Review* (1983, p. 91). I assume you've heard of that journal, Dr. Napfkavitch?"

Napfkavitch: "Yes."

Attorney: "In fact, Dr. Napfkavitch, the *AER* is one of the leading journals, if not *the* leading economics journal, is it not?"

Napfkavitch: "Yes."

Attorney: "So that, Dr. Napfkavitch, in effect you are asking the Court to accept an analysis that is in conflict with the position taken by one of the world's most highly regarded economists and published in your profession's leading journal? Isn't that so, Dr. Napfkavitch?"

Even without having to suffer the cross-examining attorney's smirk, the hitherto confident Napfkavitch would start to feel somewhat uncomfortable and perhaps wish he could roll back the clock 24 hours to a time when he was at the dentist's undergoing root canal work. Moreover, with refreshed memory his concern might heighten at the prospect of being forced to concede that, like a referee of the Fisher-McGowan paper, he too suspects IBM of having been "more profitable" than American Motors (1983, p. 82), a seemingly radical view that, he now recognizes, could strike at the very heart of his credibility.

Still further, the uneasy Napfkavitch has previously testified under direct that (a) over the past twenty years four-firm (sales and capacity) concentration levels in the market have consistently exceeded 75 percent (with comparably high Herfindahls), (b) market shares have been fairly stable, (c) the industry is characterized by high seller fixed-to-variable costs relative to other industries, (d) there has been no successful entry into the market during that period, and (e) real prices have generally increased to a greater extent than have (real) average costs, which

*Department of Management, College of Business Administration, University of Florida, Gainesville, FL 32611.

has led him to conclude that the defendant corporation probably has and exercises some market power in a market that is behaving in a "less-than-competitive manner." Thus, our hapless hero now fears that virtually *everything* upon which his conclusions have been based might be subject to a comparable attack. Does he believe that concentration ratios, say, necessarily reflect or measure market power? Of course not. Does he believe that list or published prices necessarily reflect transactions prices? Of course not. How meaningful are cost figures that include a depreciation figure that has been chosen for tax purposes and that reflects both old and new plant and equipment? How accurate are capacity data that show one seller periodically producing in excess of 100 percent of capacity? Our frenetic friend Napfkavitch is in for a long day.

Assuredly, some would maintain that the rapidly deteriorating Napfkavitch is getting just what he deserves. In the first place, he should expect and be prepared to defend his views and the data upon which they are based; and, in the second place, while his views might be defensible, the data upon which they are based are not. Our pathetic protagonist, however, is not nearly the clod that he is about to be made out to be. Rather, to paraphrase one of my occasional colleagues, having wandered into the jungle, spied some fresh elephant tracks and smelt an elephant, he is prepared to conclude that an elephant has recently wandered by.

While only God can make a tree, only accountants can make the data upon which economists are forced to rely in their antitrust analyses.[1] Given that constraint and the resulting data imperfections, all that the economist can be expected to do is to use those data to tell a story as to what has taken place in a market over time and to attempt to provide the most cogent economic explanation for that history. As sure as night follows day, "the other side" will provide its own economic expert to offer an alternative explanation. The court must then assess the opposing views. Are the court's and society's interests best served by observing that all empirical data are suspect in the sense that they do not necessarily measure what we want and presume them to measure, and that, therefore, to use those data is a "totally misleading enterprise"? I think not. Rather, it is an enterprise that must be undertaken judiciously and with an awareness of the data's possible shortcomings. Until I read the Fisher-McGowan paper, I had not considered this long-held view of mine to be especially unique nor profound, and I appreciate having been given the opportunity to discover otherwise.

[1] To be sure, economists do construct some of the data series that they employ. But, as has often been stated, an economist is a person that works with numbers but lacks the personality to be an accountant.

REFERENCE

Fisher, Franklin M. and McGowan, John J., "On the Misuse of Accounting Rates of Return to Infer Monopoly Profits," *American Economic Review*, March 1983, *73*, 82–97.

The Misuse of Accounting Rates of Return: Comment

By William F. Long and David J. Ravenscraft*

In a recent article in this *Review* (1983), Franklin Fisher and John McGowan (henceforth F-M) claim to have shown that "...there is no way in which one can look at accounting rates of return and infer anything about relative economic profitability or, a fortiori, about the presence or absence of monopoly profits" (p. 90). They then attempt to link this extremely negative conclusion exclusively to profit-concentration studies, despite more obvious and appropriate areas of application.

Aside from the questionable focus, the authors have little basis for reaching their conclusion, especially in regards to the profit-concentration issue. First, F-M do not always perform the calculations correctly. Second, they base their entire analysis on a measure of the profit rate which is not the one preferred in profit-concentration studies. Third, their examples tend to represent extreme cases; they do not reflect the typical U.S. industrial experience. Fourth, they do not demonstrate that the use of accounting rates of return leads to a positive bias in the profit-concentration relationship. And finally, they ignore substantial evidence that accounting profits do, on average, yield important insights into economic performance.

I. Analytical Errors

Fisher and McGowan's end-of-year asset accounting rates of return are incorrectly calculated. In comparing asymptotic accounting rates of return using beginning-of-year assets with those using end-of-year assets (Tables 2, 3, 5), they show the former rates as being greater than or equal to the latter, which is inconsistent with the results in their Table 1. If there is depreciation, and if the same accounting profit value is divided by the two asset values, the end-of-year accounting rate of return must be larger than the beginning-of-year accounting rate of return. Our Table 1 reproduces Panel B of F-M's Table 2, but with the correct definition of end-of-year assets.[1] There is a significant difference between the correct numbers and those reported by F-M. Therefore, their end-of-year asset results, except for their Table 1, should be discarded.

A much more serious problem is that F-M's analysis of end-of-year assets, even when correctly calculated, is still incomplete and misleading. They show in their Appendix that for continuous time, equality of the growth rate and economic rate of return assures equality of the accounting and economic rate of return. For the discrete analysis in the text, however, they show that the relationship holds for only accounting rates of return which use beginning-of-year assets as the denominator. They explicitly note that the relationship does not hold if end-of-year assets are used. The implication is that the standard practice of measuring assets as of the end of the period is incorrect. Their conclusion rests on a faulty transition from continuous analysis to discrete analysis, and on inconsistent definitions of economic rate of return, accounting rate of return, and growth rate.

The continuous time results derived in F-M's Appendix hold in discrete time for accounting profit rates defined with beginning-of-year assets as the denominator, if the growth rate and internal rate of return are defined in beginning-of-year terms. However, it also holds for accounting profit rates defined with end-of-year assets as the denominator, provided the growth rate and

*Bureau of Economics, Federal Trade Commission, Washington, D.C. 20580. The views expressed here are our own and not necessarily those of the Federal Trade Commission or any of its members. A review has been conducted to ensure that the data in this paper do not identify individual company line of business data.

[1] Only straight-line and sum-of-years' digits depreciation method results are given. F-M did not give sufficient information to permit us to distinguish among the many types of declining balance depreciation schedules.

TABLE 1—ASYMPTOTIC ACCOUNTING RATES OF RETURN (%) ON THREE VERSIONS OF THE Q-PROFILE: END-OF-YEAR ASSETS CORRECTION

	Six-Year Life (No Delay)		Seven-Year Life (One-Year Delay)		Eight-Year Life (Two-Year Delay)	
Growth Rate	Straight Line	Sum-of-Years' Digits	Straight Line	Sum-of-Years' Digits	Straight Line	Sum-of-Years' Digits
0	21.3	29.0	24.1	32.9	27.0	37.0
5	20.9	26.9	22.4	28.8	23.9	30.7
10	20.6	25.0	20.8	25.1	21.1	25.4
15	20.2	23.3	19.2	21.8	18.6	20.8
20	19.8	21.7	17.8	19.0	16.3	16.9
25	19.3	20.4	16.5	16.4	14.2	13.5
30	18.9	19.1	15.2	14.1	12.4	10.6

internal rate of return are defined in end-of-year terms. In fact, it holds for any convex combination of beginning- and end-of-year assets, subject to the requirement that the profit rate, growth rate and internal rate of return are all consistently defined.[2] Using variables defined relative to end-of-year instead of beginning-of-year, F-M's Table 2 could be recalculated to show the accounting rate of return for *end-of-year* assets equaling 15 percent for all cash flow profiles and all depreciation schedules, when the end-of-year growth rate is 15 percent. Thus, there is no a priori reason for preferring beginning- or end-of-year assets.

Fisher and McGowan's third problem is that they use 15 percent as the value of the economic rate of return, claiming that this was the average accounting rate of return for manufacturing in 1978. That is the value for the return to equity; the accounting rate of return to total assets was 7.8 percent in 1978 (FTC *Quarterly Financial Report*... *First Quarter, 1979*). If an economic rate of return of 7.8 percent is used instead of 15 percent, and the set of growth rates is centered on 7.8 percent, the maximum deviation from the economic rate of return for the accounting rate of return on beginning-of-year assets is 3.9 vs. 10.9 percentage points in F-M's Table 2 or 50 vs. 73 percent of the economic rate of return. The choice of a rate of return is important if we are trying to characterize the accuracy of accounting rate of return in the real world.

II. Alternative Profitability Measures

An analysis of the appropriate measure of profitability warrants a more extensive treatment than can be given in this brief note. However, F-M's claim that "...the economic rate of return is the only correct measure of the profit rate for purposes of economic analysis" (p. 82) requires some comment. The correct definition of profit depends on the context in which it is employed. If the analysis involves a study of investment behavior, then clearly the marginal economic rate of return is the correct profit measure.[3] It is not the preferred measure when studying monopoly power. Existing evidence suggests that the Lerner index, which can be approximated by profit/sales,[4] better reflects the degree of monopoly power.

[2] Let f_u be cash flow for an investment in the uth year of the investment's life, X_t be an arbitrary asymptotic variable in period t, r be the economic rate of return, and g be the growth rate. For beginning-of-year analysis, the definitions of r and g are given by $\Sigma_1^T(1+r)^{-u}f_u=1$ and $g=(X_{u+1}-X_u)/X_u$. For end-of-year analysis, the definitions of r and g are given by $\Sigma_1^T(1-r)^{u-1}f_u=1$ and $g=(X_{u+1}-X_u)/X_{u+1}$. The proof for the general case, of which beginning-of-year and end-of-year are special cases, is presented in our working paper (1983).

[3] Even in this context, several authors have concluded that measurement problems are not serious enough to make empirical work worthless (George Stigler, 1963; Martin Feldstein and Lawrence Summers, 1977). Others have developed procedures for correcting some of the measurement errors (Allan Young, 1975).

[4] The approximation is exact if average cost is constant. If it is not, the profit/sales ratio is a simple function of (price − marginal cost)/price and the elastic-

The basic reference on this issue is Joe Bain (1951). In describing his conceptual framework, he noted: "Average excess *profit rates on sales* should thus be higher with than without monopoly or effective oligopolistic collusion.... we arrive at the hypothesis that there will be a systematic difference in average excess *profit rates on sales* between highly concentrated oligopolies and other industries" (p. 295–96, emphasis added). In his empirical work, Bain used the ratio of accounting profits to equity, though he reported that he had also conducted his statistical tests with the ratios of excess profits to sales and accounting profits plus interest to total assets, with the same general results.

Recent contributions have expanded on this conceptual framework, including Keith Cowling and Michael Waterson (1976), Frank Gollop and Mark Roberts (1979), and Long (1982). These studies have demonstrated that profit (net of all costs) divided by sales is a performance measure which may be derived from an optimization exercise in long-run equilibrium oligopoly models that include a conjectural variable.[5] They derive estimable equations which relate this profit measure to market share, concentration, elasticity, the magnitude of the conjectural variable, and other aspects of firms and industries. These models need to incorporate an investment function and dynamic considerations (particularly entry and exit) before the issue can be fully resolved.[6] However, given that no explicit model of oligopoly derives profit rates to assets as the performance measure, the profit/sales ratio appears to be a more defensible index.

There are few explicit discussions of the appropriate profit rate in the profit-concentration literature; the best we know is Leonard Weiss (1974, p. 198–99). The issue certainly is not settled in the literature, since many studies use either equity or total assets as the denominator of the profit rate variable. Most of those cite George Stigler (1963) as the source for the preference of profit/assets over profit/sales. Their reliance on Stigler may be misplaced. He focused primarily on interactions between investment and the rate of return on assets. He defended profits divided by total assets instead of stockholders equity, but he did not directly address the use of capital instead of sales in the denominator when he considered the profit-concentration relationship.

There is an additional serious drawback with the use of a profit/assets ratio as a measure of monopoly power. If the profit/assets ratio is meant to approximate the equilibrium marginal economic rate of return on investment, then it tells us nothing about the degree of monopoly power; in equilibrium every firm, whether competitive or monopolistic, will have invested until the rate of return on the marginal investment, adjusted for risk and net of the competitive cost of capital, equals zero. If, alternatively, it is meant to approximate the *average* economic rate of return, other problems arise. The average economic rate of return *on capital* may be equal to zero in competitive industries and greater than zero in noncompetitive industries, but beyond that the magnitude of the average return on capital does not tell us anything about the *degree* of monopoly power. For example, two monopolistic industries with the same demand and constant long-run average cost curves will have the same profit/sales ratio, but their profit/assets ratio will differ to the extent

ity of the average cost curve. Thus, the profit/sales equation needs to be expanded to include determinants of the cost elasticity. For an explicit development of this approach, see Long (1982).

[5] Note profits should be net of all costs, including capital cost. A common argument for using the profit/assets ratio is that most studies have not netted out capital cost. However, if profit/sales is the theoretically correct measure, then this amounts to arguing that two wrongs make a right. The preferred method is to estimate capital cost, subtracting these estimates from profits.

[6] A model of the entry/exit process is often implicitly used in supporting the use of profit/assets ratios. However, if capital markets are competitive, the residual of revenues over all costs (including the normal return to capital) accrue to the entrepreneurship function, not to capital. It still makes sense to envision firms moving into areas where the returns are highest, but it makes no sense, from this perspective, to divide the profit residual by some measure of capital. We are indebted to David Qualls for this observation.

that their capital intensities differ. Thus, higher profit/assets ratios, even when measured correctly, do not necessarily imply greater monopoly power.

We, therefore, question the relevance of F-M's work for the profit-concentration literature, since they do not address those studies which use profit/sales as the index of performance. Furthermore, F-M ignore the difference between the marginal and average return to capital by assuming all investments have the same cash flow, an assumption that even they admit is unrealistic.

III. The Validity of F-M's Examples

Even though F-M's analysis is inaccurate and incorrectly applied, it may still be useful to examine what sort of inferences can be drawn from their theorems and examples. The theorems show that accounting profits will not equal economic profits except in special circumstances. However, for most uses, it is sufficient if accounting profits are a reasonable proxy for economic profits. The examples employed by F-M illustrate that in some cases the differences between accounting and economic profit can be fairly large. Other examples can just as easily be devised for which the differences are immaterial. The relevant questions are, which examples are more representative of the population as a whole, and whether the measurement errors introduce systematic bias in statistical studies?

Work by Thomas Stauffer (1971) sheds some light on these issues. He estimated economic profit for nine industries in which large differences between accounting and economic profits were likely. These were industries with a substantial amount of long-lived assets, *R&D* expenditures, advertising expenditures, or other special features such as capitalized sales. Despite this special selection, the correlation between accounting and economic rates of return was .79. If one could extend this work to all industries, the correlation would presumably be significantly higher. There are, of course, some industries, such as pharmaceuticals, where the difference between accounting and economic profits are large, more in line with F-M's examples. But, as Stauffer emphasizes:

> ...[T]here is little reason to expect that significant corrections would emerge for most firms, since the great majority of U.S. manufacturing industries seem to have relatively rapid inventory turnover, short gestation periods in plant construction, a comparatively low level of *R&D* or product development expenditure, and reasonably high ratios of working capital to fixed assets.... Thus, extensive corrections to indicated rates of return should be the exception, rather than the rule. [p. V–10]

F-M's examples, therefore, do not appear to represent the typical industry.

Fisher and McGowan's use of an accelerated depreciation schedule in their examples creates exaggerated accounting-economic rate of return differences. In all of their examples except Table 2, they employ a sum-of-years' digits depreciation schedule. Using the 1975 line of business survey of 472 large manufacturing companies, we calculated that approximately 80 percent of assets were depreciated with the straight-line procedure. Only about 9 percent use sum-of-years' digits. The use of straight-line depreciation in all of the examples would therefore be more appropriate if F-M wish to claim their examples are representative.[7] The depreciation method selected is important, as can be seen in F-M's Table 2. The extent to which the accounting rate differs from the economic rate is substantially smaller for the straight-line method than the accelerated depreciation schedules.

The fundamental problem is that F-M try to reach general conclusions about statistical relationships through examples. Such an attempt is fundamentally flawed, since the examples may only reflect extremes. The inaccuracy of this approach can be illustrated from other aspects of the profit measurement problem. Line of business (*LB*)

[7] In its measurement of national income and related macroeconomic variables, the Bureau of Economic Analysis of the Department of Commerce converts all depreciation to the straight-line method. For a justification, see Young (pp. 15, 35).

profits may be distorted because of common cost allocations or nonmarket transfer prices. George Benston (1979) and William Breit and Kenneth Elzinga (1981) illustrate through examples that in some cases these distortions can be quite large. They, therefore, conclude that the *LB* data are misleading. Although the *LB* data set does contain some profits which are significantly affected by these problems, work by Ravenscraft (1981) and Long et al. (1983, pp. 45–63) shows that large distortions are atypical. The correlation between *LB* profits as reported by the companies and *LB* profits based on a market-allocation procedure is approximately .89. Similarly, reported *LB* profits, for which 50 percent of the transfers are valued at nonmarket prices, and *LB* profits, where all transfers are valued at market prices, also have a correlation of approximately .89.

IV. The Usefulness of Accounting Profit Data in Structure-Performance Analysis

The required accuracy of accounting profits is dependent on the context in which they are used. If a single accounting number is employed as evidence in an antitrust case, then certainly the accuracy of that number and not the typical accounting number needs to be ascertained. It is in this context that the F-M paper originated. However, they claim their analysis is relevant more generally to the profit-concentration literature without providing a justification for this extension. In particular, F-M never demonstrate (or even claim) that the use of accounting rates of return tends to overestimate economic rates of return in concentrated industries relative to unconcentrated ones, which they must do to show that the accounting-economic-profit divergence leads to a positive bias in the concentration-accounting profits relationship. If this divergence represents only random noise, then the statistical relationship between profits and concentration must be stronger than previous work indicates, because it prevails over significantly more noise than previously assumed.

Using the FTC's *LB* data for 1975, we have developed some indirect evidence that the accounting-economic-profit divergence does not significantly effect the qualitative conclusions of structure-profits regressions, even though it may introduce distortions for some individual profit numbers. As a first step, we calculated ordinary least squares (*OLS*) regression statistics for a leading equation in Ravenscraft (1983, p. 26), using profits/sales and profits/end-of-year assets as dependent variables, and using the same independent variables as he used.[8] The profit/sales regression and the profit/assets regression yield similar structure-performance inferences, with respect to most of the key variables, a result which is consistent with the findings of Bain and other researchers. In addition, the strongest statistical results arose in the profit/sales regression, which lends support to the choice of profits/sales over profits/assets as the dependent variable in such regressions.

The corrected F-M examples point to the potential for a large difference between profits as a ratio to beginning-of-year and end-of-year assets, when there is a substantial accounting-economic profit divergence. Therefore, structure-profit regressions using profits/beginning-of-year assets and profits/end-of-year assets should yield different statistical inferences, if the accounting-economic profit divergence results in a significant bias. Implicit in their analysis is the expectation that midyear assets should give intermediate results. To test these hypotheses, we recalculated the profits/assets equation, but with midyear assets and beginning-of-year assets as the denominator. Qualitative conclusions about individual independent variables for the three equations are almost identical. Therefore, there is little indication of a significant bias. The R^2 with either midyear or beginning-of-year assets in the denominator is substantially higher than R^2 with end-of-year assets, but those two variants yield virtually indistinguishable results.

A third sensitivity test also indicates that the structure-profit results are generally not as biased as F-M suggest. If accounting depreciation corresponds to economic depreci-

[8] A detailed description of the regression results appears in our working paper (1983).

ation, then accounting and economic profits are equal. Therefore, it is the divergence between accounting and economic depreciation that causes the accounting and economic profit divergence. If depreciation is a relatively unimportant part of profitability, then the difference between accounting and economic profits should be small. To test for the impact of depreciation, the equation with profits/sales as the dependent variable was reestimated using profits before depreciation in the numerator. The statistical significance (using a 5 percent cutoff) of five of the twenty-three independent variables change as a result of the exclusion of depreciation from profits. These include minimum efficient scale, supplier concentration, industry vertical integration, industry advertising, and industry *R&D*. Therefore, mismeasurement of accounting profits does present some potential for distorting certain structure-profit results. However, F-M appear to be incorrect in their implication that the profit-concentration relationship is one of the results affected, since inferences concerning market share and the concentration ratio were not effected. This exercise was repeated using profit + depreciation/assets + accumulated depreciation as the dependent variable. The changes were even less significant.

V. General Evidence of Accounting Profits Usefulness

Fisher and McGowan ignore a substantial amount of evidence which demonstrates the usefulness of accounting profit data. For example, a sizable literature exists relating accounting profit to stock market values. After an extensive review of this literature, William Beaver (1981) concluded that almost all studies show a significant positive relationship between accounting earnings changes and stock market price changes, and that prices behave as if accounting earnings data "...are a potentially important source of information, but only one of many sources" (p. 118). Assuming that the stock market reflects knowledge of economic profits, accounting profits must do the same, at least to some degree, if investors consider them useful.

Accounting profit data are used to evaluate numerous economic issues besides questions in industrial organization. F-M have little justification for focusing solely on their use in evaluating monopoly. Many studies have used accounting profits to demonstrate the efficiency of large firms. Why not title the paper "On the Misuse of Accounting Rates of Return to Infer Efficiency?" The investment-profit literature is just as vast and important in terms of public policy as the profit-concentration literature, yet F-M did not even reference this literature, despite the fact that rate of return is the central concept in the investment literature. The growth and productivity literature implicitly assumes depreciation and assets are correctly measured. Even many basic measurements in macroeconomics, such as *GNP*, are dependent on accounting profit data. Therefore, the implication of F-M's work, if correct, is that most of applied economics is misguided.

The broad use of accounting profit data in the private sector suggests that F-M's general conclusions about the uselessness of the data must be wrong. They are certainly valuable by a simple market test—private firms spend vast resources collecting and analyzing them. A large number of commercial information services (Dun and Bradstreet, Moodys, Value Line, Standard and Poors, COMPUSTAT, etc.) supply data on accounting profit rates and/or comparative analyses across firms or industries. Given the amount spent in the private sector on analyses of accounting profit data, a substantial market failure is required to explain such an occurrence if the data are valueless.

VI. Conclusion

The flaws detailed above substantially limit the applicability of F-M's work. The evidence they presented does not support the conclusion that accounting profit figures are meaningless. The paper simply implies that individual accounting profit numbers can under certain circumstances deviate significantly from economic profits. However, there is no evidence that large deviations exist on average. Fisher and McGowan are equally wrong in their contention that the profit-concentration literature is a "misleading enterprise." They give no indication as to how accounting mismeasurement biases the profit-concentration relationship. The evidence presented in this comment suggests a bias does not exist.

REFERENCES

Bain, Joe. S., "Relation of Profit Rate to Industry Concentration: American Manufacturing, 1936–1940," *Quarterly Journal of Economics*, August 1951, *65*, 293–324.

Beaver, William H., *Financial Reporting: An Accounting Revolution*, Englewood Cliffs: Prentice-Hall, 1981.

Benston, George J., "The FTC's Line of Business Program: A Benefit-Cost Analysis," in Harvey Goldschmid, ed., *Business Disclosure: Government's Need to Know*, New York: McGraw-Hill, 1979, 58–118.

Breit, William and Elzinga, Kenneth G., "Information for Antitrust and Business Activity: Line of Business Reporting," in Kenneth W. Clarkson and Timothy J. Muris, eds., *The Federal Trade Commission since 1970: Economic Regulation and Bureaucratic Behavior*, Cambridge: Press Syndicate of the University of Cambridge, 1981, 98–120.

Cowling, Keith and Waterson, Michael, "Price-Cost Margins and Market Structure," *Economica*, August 1976, *43*, 267–74.

Feldstein, Martin S. and Summers, Lawrence H., "Is the Profit Rate Falling?," *Brookings Papers on Economic Activity*, 1:1977, 211–28.

Fisher, Franklin M. and McGowan, John J., "On the Misuse of Accounting Rates of Return to Infer Monopoly Profits," *American Economic Review*, March 1983, *73*, 82–97.

Gollop, Frank M. and Roberts, Mark J., "Firm Interdependence in Oligopolistic Markets," *Journal of Econometrics*, August 1979, *10*, 313–31.

Long, William F., "Market Share, Concentration and Profits: Intra-industry and Inter-industry Evidence," unpublished paper, Federal Trade Commission, 1982.

______, **Lean, David F., Ravenscraft, David J., and Wagner, Curtis L. III**, *Benefits and Costs of the Federal Trade Commission's Line of Business Program*, Washington: Federal Trade Commission, 1983.

______ **and Ravenscraft, David J.**, "The Usefulness of Accounting Profit Data: A Comment on Fisher and McGowan," Working Paper No. 94, Federal Trade Commission, June 1983.

Ravenscraft, David J., "Structure-Profit Relationships at the Line of Business and Industry Level," *Review of Economics and Statistics*, February 1983, *65*, 22–31.

______, "Transfer Pricing and Profitability," unpublished paper, Federal Trade Commission, 1981.

Stauffer, Thomas R., "Measurement of Corporate Rates of Return," unpublished doctoral dissertation, Harvard University, 1971.

Stigler, George J., *Capital and Rates of Return in Manufacturing Industries*, Princeton: National Bureau of Economic Research, 1963.

Weiss, Leonard W., "The Concentration-Profits Relationship and Antitrust," in Harvey J. Goldschmid et al., eds., *Industrial Concentration: the New Learning*, Boston: Little, Brown and Company, 1974, 184–232.

Young, Allan H., "New Estimates of Capital Consumption Allowances in the Benchmark Revision of GNP," *Survey of Current Business*, October 1975, *55*, 14–16, 35.

U.S. Federal Trade Commission, *Quarterly Financial Report for Manufacturing, Mining and Trade Corporations: First Quarter, 1979*, Washington, 1980.

______, *Annual Line of Business Report–1975*, Washington, 1981.

The Misuse of Accounting Rates of Return: Comment

By STEPHEN MARTIN*

> It is better to light one candle than curse the darkness.
>
> *Motto of the Christopher Society*

In a recent paper (1983), Franklin Fisher and John McGowan argue that the large empirical literature that purports to examine the relationship between market concentration and profitability in fact does not do so, and that such exercises with accounting measures of profitability are meaningless.

These conclusions are based on the following series of propositions:

1) empirical investigations of the relationship between concentration and profitability "uniformly" measure profitability as a rate of return on assets or stockholders' equity;

2) the discount rate that makes the present value of an income stream equal to the expenditure that generates the income stream is the only measure of profitability which is "correct" for economic analysis;

3) (as shown by a series of examples) accounting measures of the rate of return on assets are very poor proxies for this "correct" rate of return.

I wish to make the following points: (a) proposition 1) is inaccurate as a description of the literature that investigates the relationship between concentration and profitability; (b) proposition 2), that what Fisher and McGowan label "the economic rate of return" is the only correct measure of profitability for economic analysis, is unconvincing, since a quite different measure of profitability emerges from familiar formal models of firm behavior; (c) the Fisher-McGowan examples illustrate a property of accounting measures of assets that is well known to students of industrial organization. I will discuss each of these points in turn.

I. Empirical Studies of Profit and Concentration

Fisher and McGowan indicate:

> The large volume of research investigating the profits-concentration relationship uniformly relies on accounting rates of return, such as the ratio of reported profits to total assets or to stockholders' equity as the measure of profitability to be related to concentration. [p. 82]

As a description of the literature reporting studies of the relationship between concentration and profitability, this is simply incorrect. A large number of studies, possibly a majority, measure profitability as a rate of return on sales. This includes all studies that use the well-known "price-cost margin" computed from Census of Manufactures data (see my 1979 article, p. 474). Many of these studies are discussed in the literature surveys that are cited by Fisher and McGowan (Leonard Weiss, 1974; F. M. Scherer, 1980). Weiss (p. 199) suggests that such a measure is superior to other measures of profitability; so does Scherer (p. 269), who describes the rate of return on stockholders' equity and the rate of return on capital as "second-best" in comparison with the rate of return on sales.

II. Measuring Profitability for Economic Analysis

Fisher and McGowan state:

> The economic rate of return on an investment is...that discount rate that equates the present value of its expected net revenue stream to its initial outlay. ...it is clear that it is the economic rate of return that is equalized within an industry in long-run industry competitive equilibrium and (after adjustment for risk) equalized everywhere

*Associate Professor, Department of Economics, Michigan State University, East Lansing, MI 48824-1038. I am grateful to M. A. Adelman and Franklin Fisher for useful comments; responsibility for errors is my own.

> in a competitive economy in long-run equilibrium. It is an economic rate of return (after risk adjustment) above the cost of capital that promotes expansion under competition and is produced by output restriction under monopoly. *Thus, the economic rate of return is the only correct measure of the profit rate for purposes of economic analysis.*
> [p. 82, emphasis added]

The conclusion that what Fisher and McGowan call the economic rate of return is the only correct measure of the profit rate for purposes of economic analysis is thus based on an appeal to the economic theory of the behavior of the profit-maximizing firm. It is, however, well known that another measure of profitability arises naturally in formal models of profit-maximizing firm behavior: the Lerner index of monopoly power, the price-marginal cost margin.

Many studies of the relation between concentration and profitability have used models of the price-cost margin.[1] It manifests itself not only in models of output determination under conditions of market power, but also in models of various sorts of conduct (such as advertising or research and development).[2]

It is doubtful whether any measure of profitability can be unambiguously identified as "correct," to the exclusion of all others, for purposes of economic analysis. Fisher and McGowan's discussion of what they call "the economic rate of return" does not establish that measures of profitability based on the Lerner index are inappropriate for economic analysis.

At this point it is convenient to formally derive a version of the Lerner index. Consider a firm that combines the services $L(t)$ of labor and $K(t)$ of capital according to a continuous, twice differentiable production function $Q = F(L, K)$.[3] Output is sold at a price given by a continuous, twice differentiable inverse demand function $p(Q)$.[4] The services of labor are hired in a competitive labor market at wage $w(t)$. Capital is purchased at price $p^k(t)$. Capital stock depreciates at rate $\delta(t)$,[5] so that investment at time t is

$$I(t) = \dot{K}(t) + \delta(t)K(t), \tag{1}$$

where the dot indicates the time derivative.

The firm acts to maximize the present discounted value of net cash flow,

$$\pi = \int_{t=0}^{\infty} e^{-rt}\{p[F(L(t),K(t))]F(L(t), K(t)) - w(t)L(t) - p^k(t)I(t)\}\,dt. \tag{2}$$

First-order necessary conditions for profit maximization follow by substituting (1) into (2) and applying Euler's equation from the calculus of variations. They are

$$Qp'(Q)F_L(L,K) + p(Q)F_L(L,K) = w \tag{3}$$

$$Qp'(Q)F_K(L,K) + p(Q)F_K(L,K) = (r+\delta)p^k - \dot{p}^k = \lambda p^k \tag{4}$$

where

$$\lambda = (r+\delta) - (\dot{p}^k/p^k) \tag{5}$$

is the shadow rental of the quantity of capital

[1] For example, Joe Bain (1956, pp. 7; 190–91); Norman Collins and Lee Preston (1970, p. 10); Stephen Rhoades and Joe Cleaver (1973, p. 91); Keith Cowling and Michael Waterson (1976); see also S. J. Liebowitz (1982, p. 231, fn. 1).

[2] Richard Schmalensee (1972, pp. 20–43); Douglas Needham (1975); John Cubbin (1981).

[3] For references to the extensive literature on aggregation, see Robert Solow (1956) or Fisher (1969). The standard neoclassical model is worth investigating, not as an exact description of reality but as a useful approximation to reality. For a specific discussion of aggregation and empirical studies of industrial organization, see my paper with David Ravenscraft (1982).

[4] The firm may be a pure monopolist or a producer in a monopolistically competitive industry. The demand function and the production function may be made to depend on time without altering the nature of the results.

[5] As δ is allowed to vary over time, this is not a "Santa Claus" case (Fisher and McGowan, p. 92). It is possible to endogenize the rate of depreciation (as a function of the intensity of use of capital) without altering the nature of the results.

that may be purchased for a dollar (and functional dependence on time has been suppressed for compactness).

Equations (3) and (4) may be rewritten

$$(6) \qquad (p - w/F_L)/p = 1/\varepsilon_{Qp}$$

$$(7) \qquad (p - \lambda p^k/F_K)/p = 1/\varepsilon_{Qp}$$

where ε_{Qp} is the price elasticity of demand. Of course, equations (6) and (7) are just two different ways of writing the Lerner formula, since marginal cost is

$$(8) \qquad MC = w/F_L = \lambda p^k/F_K.$$

Equations (6) and (7) thus imply

$$(9) \qquad (p - MC)/p = 1/\varepsilon_{Qp}$$

and the Lerner index of monopoly power, the price-marginal cost margin, has emerged from a formal, dynamic, intertemporal model of profit-maximizing firm behavior.

It is important to note that the optimal conditions for factor employment at time t (equations (6) and (7)) and the equivalent Lerner index depend only on values that are known at time t, a property that Kenneth Arrow terms "myopia:

> ...[P]erhaps the most striking feature of the optimal policy is its independence of future movements of the profit function. This function, it must be remembered, incorporates all knowledge of market conditions both for the selling of the product and for the purchasing of inputs; it also incorporates all aspects of technology other than depreciation of equipment. In particular, the future shifting of technological knowledge plays no role in present investment decisions.
>
> The myopic property of the optimal capital policy implies a considerable economy of information needs in the firm's decision making process, perfectly comparable to the use of the price system for decentralization.
>
> Until very recently the myopic property was largely unremarked in the literature. Indeed, the usual formulation, for example, Keynes's use of the marginal efficiency of capital,...requires comparison of the present value of all future returns for a given investment with the investment cost. This procedure is not unambiguous...its most significant defect is to concentrate attention on the choice between undertaking an investment and not undertaking it at all, whereas the myopic rule is based on comparison between undertaking the investment now and postponing it for a short period.
>
> [1964, pp 27–28]

When factor markets work—when the price system allows decentralization in factor markets—the myopic property similarly allows a considerable economy of information in the assessment of profitability, by use of measures based on the Lerner index rather than present value calculations.

Such measures of profitability will be suitable for samples of firms or industries that employ relatively nonspecific, tradable capital assets (such as wholesale or retail trade; see my forthcoming article and Bruce Marion et al., 1979).[6] Such measures of profitability will be appropriate for samples drawn from populations that employ specialized, imperfectly tradable capital assets, if the degree of specificity of assets is roughly constant over the sample (Blake Imel et al., 1972). Such measures of profitability will be appropriate for broad cross-section samples if one controls for variations across the sample in the nature of markets for capital assets (and it can be argued that conventional measures of absolute capital requirements and entry conditions do this, in an imperfect way).

The myopic property of the Lerner index also reflects the fundamental differences between the conventional neoclassical view of the production process, that underlies the model presented here, and the view of the production process that is implicit in the Fisher-McGowan examples. The firm is

[6] Alternatively, one may say that sunk costs should be small and fixed assets traded in markets which work well; see William Baumol et al. (1982, pp. 280–82 and fn. 2).

modeled here as an ongoing concern, acquiring fixed and variable factors to produce output that is marketed on a continuing basis; decisions are made by evaluating the consequences of marginal changes. The examples presented by Fisher and McGowan involve what might be called "oilfield production": as asset is acquired, a project is undertaken, the project yields a time stream of returns, the asset is used up, the project is ended.[7] Ongoing firms in the Fisher-McGowan examples are simply collections of such assets, that are not traded once a project commences.

III. The Lerner Index and Accounting Measurements

Fisher and McGowan cannot establish that what they call the economic rate of return is the unique correct measure of profitability for purposes of economic analysis. They simply fail to discuss the large portion of the empirical literature relating concentration to measures of profitability based on the Lerner index. Their conclusions concerning the literature that relates concentration and profitability are thus not established by their arguments.

However; their examples do illustrate an important property of accounting data, one that has implications for the use of the Lerner index. I show this by specializing equation (9) and obtaining an expression for the Lerner index that can be related to empirical studies employing rates of return on sales as measures of profitability.

Suppose the production function exhibits constant returns to scale; as noted by William Baumol et al. (p. 33), this is the leading empirical case. Under constant returns to scale, marginal cost and average cost are the same. Formally, from (8),

$$(10)\quad wL+\lambda p^kK = MC(LF_L+KF_K) = MC(Q),$$

so that

$$(11)\quad MC=(wL+\lambda p^kK)/Q=AC,$$

and the Lerner index (9) may be rewritten as

$$(12)\quad (pQ-wL)/pQ=1/\varepsilon_{Qp}+\lambda p^kK/pQ.$$

Without loss of generality, L may be interpreted at this point as a vector of variable factors, with w a conformable vector of factor prices. The left-hand side of (12) is then the margin of revenue over the cost of variable inputs, as a fraction of revenue (or equivalently the margin of price over average variable cost, as a fraction of price). This clearly corresponds to the widely used "price-cost margin" computed from Census of Manufactures data,[8] and to profit rates on sales computed from other sources.

Most simply, it may then be argued that the price elasticity of demand for the product of a single firm will be a function of industry characteristics (including but not limited to market concentration and entry conditions); more formal approaches are possible (for example, Keith Cowling and Michael Waterson). Aggregation to the industry level raises the well-known industry boundary problem—the classification of firms or divisions of firms into industries—but equation (12) clearly provides a framework that encompasses many empirical studies of profitability at the firm and industry level. Such studies should, of course, control for differences between average and marginal cost; conventional measures of minimum efficient scale and the cost disadvantage of smaller firms serve this role. It is generally

[7] As noted by Fisher and McGowan (p. 84, fn. 9), it is not the wearing out of assets that is critical. I use the term "oilfield production" with reference to Richard Mancke (1974), who runs simulations based on assumptions similar to those of the Fisher-McGowan examples.

[8] For specific discussions of the price-cost margin as computed from Census of Manufactures data, see Weiss (p. 199) or Scherer (pp. 271–72). Liebowitz criticizes the census price-cost margin by comparison with Internal Revenue Service data. Scherer identifies two major problems with IRS data: the assignment of entire firms to a single industry (p. 270) and the impact of accounting rules that are followed for tax purposes only (p. 272). Liebowitz corrects for the first problem (pp. 238–39, fn. 22); he recognizes the second (p. 238, fn. 21) and assumes it can be ignored. There is no reason to think that this is the case; his results can be interpreted as confirming the suitability of census data.

recognized that a capital-sales ratio should be included as an explanatory variable when profitability is measured as a rate of return on sales, and (12) provides a formal rationale for this.

The force of the Fisher-McGowan examples, applied to (12), is that accounting measures of the value of the capital stock are likely to be poor measures of the economic value of such assets, so that the capital-sales ratio, the second right-hand side term in (12), will be subject to serious measurement error.

It should first be noted that although the Fisher-McGowan examples make this point with great clarity, it is not new. It is discussed by Scherer (pp. 272–73); it is discussed by Weiss (pp. 196–97); it is discussed and specifically addressed by studies which employ stock market measures of asset value (James Bothwell and Theodore Keeler, 1976; Timothy Sullivan, 1977; Stavros Thomadakis, 1977).[9]

Somewhat more generally, the Fisher-McGowan examples suggest that since accounting measures of the value of capital are likely to be flawed, accounting techniques should themselves be the subject of analysis. As noted by Nicholas Gonedes and Nicholas Dopuch (1979, p. 407),[10] this is only possible where data sets include information on the nature of the accounting conventions used to record asset values. The only major cross-section data set that includes this information is the Federal Trade Commission's Line of Business data set. Two studies that examine the robustness of results of concentration-profitability studies to the use of alternative accounting conventions and alternative definitions of capital stock find that such results are robust.[11]

IV. Conclusion

The price-average cost margin or rate of return on sales is a measure of profitability which may be used for economic analysis. Fisher and McGowan have demonstrated the well-known point that accounting measures of capital intensity are likely to be inaccurate. This should be, and has been, considered in carrying out empirical studies of the concentration-profitability relationship. The literature that relates concentration to rates of return on sales constitutes a well formulated body of empirical economic research, and examination of absolute or relative price-cost margins to draw conclusions about market power can be expected to yield accurate information about structure-conduct-performance relationships.

[9] As noted by Thomadakis (p. 181, fn. 1), this measurement error is a serious problem only if systematically related to market structure.

[10] Gonedes and Dopuch are critical of studies that criticize accounting measures of income with reference to "true" or "ideal" concepts of income (pp. 384–85). They assert that the fundamental problem of accounting measurement arises in the context of incomplete or imperfect markets (p. 392, fn. 10).

[11] William Long (1981); my 1981 manuscript. The Long paper employs what Gonedes and Dopuch call a recomputation technique. The Fisher-McGowan paper is an example of what Gonedes and Dopuch call the simulation approach. As Gonedes and Dopuch note, "Neither approach dominates another in terms of insights provided" (p. 400).

REFERENCES

Arrow, Kenneth J., "Optimal Capital Policy, the Cost of Capital, and Myopic Decision Rules," *Annals of the Institute of Statistical Mathematics*, 1964, *16*, 21–30.

Bain, Joe S., *Barriers to New Competition*, Cambridge: Harvard University Press, 1956.

Baumol, William J., Panzar, John C. and Willig, Robert D., *Contestible Markets and the Theory of Industry Structure*, New York: Harcourt Brace Jovanovich, 1982.

Bothwell, James L. and Keeler, Theodore E., "Profits, Market Structure and Portfolio Risk," in Robert T. Masson and P. David Qualls, eds., *Essays on Industrial Organization in Honor of Joe S. Bain*, Cambridge: Ballinger Publishing Company, 1976, 71–88.

Collins, Norman R. and Preston, Lee E., *Concentration and Price-Cost Margins in Manufacturing Industries*, Berkeley: University of California Press, 1970.

Cowling, Keith and Waterson, Michael, "Price-Cost Margins and Market Structure," *Economica*, August 1976, *43*, 267–74.

Cubbin, John, "Advertising and the Theory of Entry Barriers," *Economica*, August 1981, *48*, 289–98.

Fisher, Franklin M., "The Existence of Aggregate Production Functions," *Econometrica*, October 1969, *37*, 553–57.

______ **and McGowan, John J.,** "On the Misuse of Accounting Rates of Return to Infer Monopoly Profits," *American Economic Review* March 1983, *73*, 82–97.

Gonedes, Nicholas J. and Dopuch, Nicholas, "Economic Analyses and Accounting Techniques: Perspective and Proposals," *Journal of Accounting Research*, Autumn 1979, *17*, 384–410.

Imel, Blake, Behr, Michael R. and Helmberger, Peter G., *Market Structure and Performance*, Lexington: D. C. Heath, 1972.

Liebowitz, S. J., "What Do Census Price-Cost Margins Measure," *Journal of Law and Economics*, October 1982, *25*, 231–46.

Long, William F., "Impact of Alternative Allocation Procedures on Econometric Studies of Structure and Performance," manuscript, Line of Business Program, Federal Trade Commission, July 1981.

Mancke, Richard B., "Causes of Interfirm Profitability Differences: A New Interpretation of the Evidence," *Quarterly Journal of Economics*, May 1974, *88*, 181–93.

Marion et al., Bruce W., "The Price and Profit Performance of Leading Food Chains," *American Journal of Agricultural Economics*, August 1979, *61*, 420–33.

Martin, Stephen, "Entry Barriers, Concentration, and Profits," *Southern Economic Journal*, October 1979, *46*, 471–88.

______, "Modeling Profitability at the Line of Business Level," manuscript, Line of Business Program, Federal Trade Commission, August 1981.

______, "Structure and Performance of U.S. Wholesale Trade," *Managerial and Decision Economics*, forthcoming, 1984.

______ **and Ravenscraft, David J.,** "Aggregation and Studies of Industrial Profitability," *Economics Letters*, No. 1; 2, 1982, *10*, 161–65.

Needham, Douglas, "Market Structure and Firms' R&D Behavior," *Journal of Industrial Economics*, June 1975, *23*, 241–55.

Rhoades, Stephen A. and Cleaver, Joe. M., "The Nature of the Concentration-Price/Cost Margin Relationship for 352 Manufacturing Industries: 1967," *Southern Economic Journal*, July 1973, *40*, 90–102.

Scherer, F. M., *Industrial Market Structure and Economic Performance*, Chicago: Rand McNally, 1980.

Schmalensee, Richard, *The Economics of Advertising*, Amsterdam: North-Holland, 1972.

Solow, Robert M., "The Production Function and the Theory of Capital," *Review of Economic Studies*, No. 2, 1956, *23*, 101–8.

Sullivan, Timothy G., "A Note on Market Power and Returns to Stockholders," *Review of Economics and Statistics*, February 1977, *59*, 108–13.

Thomadakis, Stavros B., "A Value-Based Test of Profitability and Market Structure," *Review of Economics and Statistics*, May 1977, *59*, 179–85.

Weiss, Leonard W., "The Concentration-Profits Relationship and Antitrust," in Harvey J. Goldschmid et al., eds., *Industrial Concentration: the New Learning*, Boston: Little, Brown, and Co., 1974, 184–233.

The Misuse of Accounting Rates of Return: Comment

By MICHAEL F. VAN BREDA*

Franklin Fisher and the late John McGowan are to be commended for drawing the attention of economists to the very real measurement problems that exist when accounting rates of return are used in place of economic rates of return. My own research in this area confirms the general tenor of their paper; however, the cause is not completely hopeless as they will agree.

They couch their findings in very careful terms stating that "the accounting rate of return depends *crucially* on the time shape of benefits, and the effect of growth on the accounting rate of return also depends on that time shape" (p. 84). Since this time shape is unknown to outsiders, no relationship between the accounting rate of return and the economic rate of return can be established by them. The point they make is well taken and needs to be made. What they fail to stress though is that *if* the "time shape" is known, *then* a relationship can be shown to exist.

In my paper on this topic (1981a), I demonstrated that if the time shape of real benefits from a single asset were constant (i.e., the one-hoss shay example) and the company were in a state of constant growth, then a series of graphs could be drawn linking the economic rate of return through the accounting book life and the growth of the assets to the accounting rate of return. In the example I treated, the accounting rate of return exceeds the economic rate of return whenever the growth rate falls below the economic rate. The divergence increases with increasing book life and also with the rate of depreciation. Similar sorts of generalizations can be made for other time shapes although the direction of the divergence varies from case to case as Fisher and McGowan point out—a point I failed to make.

There can be no question that unless the time shape be given, no systematic relationship can be established. On the other hand, it is my personal experience that managers can and do estimate time shapes of benefits. How else would they do capital budgeting? Management can therefore establish "equilibrium"-type hurdle rates for divisions that are based on systematically adjusted economic rates. In fact, the model to which I alluded was used by a Fortune 500 firm to analyze their hurdle rate structure. To the extent that these time shapes are available to outsiders, such as researchers, it should not be impossible for them to do the same.

In dealing with this problem with management, I have since come to believe that there is a further problem related to accounting rates that needs addressing, and in some sense is even more basic than those touched on above. All our analyses of accounting rates of return are based, by necessity, on stationary states or steady-state growth scenarios—what Fisher and McGowan refer to as the "most favorable cases"—and what I shall refer to as equilibrium states in the sense that the parameters in those states are constant or "at rest." In practice, such states are rarely encountered, making most of our results in the area, including my own, of slightly dubious value.

The problem can be finessed in one direction. When a company is attempting to set a hurdle rate for a division, it can look upon this rate as an equilibrium rate and argue that when the division reaches the desired steady state growth, then it should be earning the following accounting rate of return. One cannot work backwards, though, because the accounting rates one is seeking to convert back to economic rates were not generated in a state of dynamic equilibrium in the sense in which we are using that term here. This was the insoluble dilemma we faced in our real world analysis.

*Edwin L. Cox School of Business, Southern Methodist University, Dallas, TX 75275.

This situation is not unique to accounting. Economists have little to say in any general sense about prices in a state of disequilibrium. The difficulty the economist has is compounded by the often overlooked dynamics of the accounting system itself. One example of this accounting dynamic is that we will be well into the twenty-first century before the effects of 1981's inflation have wound their way out of the balance sheets of U.S. companies. In other words, it will be twenty years before the accounting system will move into anything like a steady state—presuming that in the meantime no further inflation is experienced.

Some of the implications of the dynamics of accounting rates of return are described in my monograph (1981b) where I argue that accounting rates should be treated as the output from an accounting "filter," the input to which are economic events. If we can establish satisfactory dynamic models of this input process, then it should not be difficult to model the output. Given that management and the public at large, almost certainly, and despite all our strictures, will continue to measure the success of a company by its accounting rate of return, it seems to me that continued research into the relationship between accounting and economic rates of return is very necessary.

REFERENCES

Fisher, Franklin M. and McGowan, John J., "On the Misuse of Accounting Rates of Return to Infer Monopoly Profits," *American Economic Review*, March 1983, *73*, 82–97.

Van Breda, Michael F., (1981a) "Accounting Rates of Return under Inflation," *Sloan Management Review*, Summer 1981, *22*, 15–28.

______, (1981b) *The Prediction of Corporate Earnings*, Ann Arbor: UMI Research Press, 1981.

The Misuse of Accounting Rates of Return: Reply

By FRANKLIN M. FISHER*

After he had made his remark about the Emperor's new clothes, the boy was quite surprised at the reaction. One commentator pointed out how hard the boy had made it for established experts on imperial robes to testify about what they just *knew* to be true. His comments were at least amusing, but others tended to be more strident. Indeed, some designers and fashion critics (particularly those involved in line-of-royalty fashion reporting) accused the boy of trying to destroy not only the clothing business, but all of business generally. They pointed out that the boy's model emperor was unrepresentative of true emperors and, anyway, emperors and clothes-wearing tended to be correlated. Still others said that, while it was true that the Emperor had no clothes, he did carry a stick, and, under some assumptions, sticks could be sceptres.

While a few people did say that they agreed with the boy, they were mostly people who had already remarked on the state of imperial undress. Indeed, since a number of people had said much the same thing in the past, the boy wondered why he should have provoked such strong reactions. Perhaps, he thought, it was because he had been unduly blunt in referring to His Majesty as "that naked jaybird" or because he had published his findings in the very prestigious *Court Journal.*

Moral: *Don't interfere with fairy tales if you want to live happily ever after.*

I. What Did We Say?

Judging from some of the comments I have received, only some of which are published above, you would think that John McGowan and I had defaced a national monument. We have been accused of claiming that all accounting data are useless, of making it difficult for expert economists to testify about monopoly, and even of implying that "most of applied economics is misguided." To put things in perspective, let's examine what we did say.

McGowan and I pointed out that the profit rate about which economic theory speaks is the internal or economic rate of return. It is that rate, if any, which provides the signal for the entry or exit of firms and resources. Hence, studies which use profit rates as though they were the objects analyzed in the theory must use profit rates which measure the economic rate of return if those studies are to be worth anything at all. This is particularly true of studies either of individual firms or of cross-firm-or-industry comparisons which seek to identify high rates of return with monopoly power.

Such studies use accounting rates of return, defined as current profits divided by either total capitalization or the value of stockholders' equity.[1] For convenience, we concentrated on the first alternative, although our qualitative conclusions applied to both. We ignored well-known difficulties such as the treatment of intangibles in order to concentrate on the conceptual problem involved.

That problem is as follows. The numerator of the accounting rate of return in question is current profits; those profits are the consequence of investment decisions made in the past. On the other hand, the denominator is total capitalization, but some of the firm's capital will generally have been put in place relatively recently in the expectation of a profit stream much of which is still in the future. While the economic rate of return is the magnitude that properly relates a stream of profits to the investments that produce it, the accounting rate of return does not. By

*Professor of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139. I am grateful to Joen E. Greenwood, Paul Joskow, and Robert Larner for comments, but retain responsibility for error.

[1] Our paper did not discuss the use of profits divided by sales. I take this up below.

relating *current* profits to *current* capitalization, the accounting rate of return fatally scrambles up the timing.

Moreover, this defect is not something that can be corrected by averaging, nor is it merely a start-up problem. It persists even in steady-state growth. Unless the firm values its assets in the particular way long ago pointed out by Harold Hotelling, its accounting rate of return will not equal the economic rate of return. Further, that particular valuation method is totally impractical. For firms to use it would sometimes require taking negative depreciation; for observers to do so would require knowing the economic rate of return, so that computing the accounting rate would be pointless.

Without the use of Hotelling valuation, the accounting rate of return is not very informative, to say the least. In particular, even in the most unrealistically favorable case, that of exponential growth, the most that can be said is that the accounting and economic rates of return are always on the same side of the growth rate—a fact that is not very helpful. In a series of computer-generated examples, McGowan and I showed that the effects involved could be very large indeed, so that firms with the same economic rate of return could have widely different accounting rates of return. We also showed that ranking of firms by accounting rates could invert their ranking by economic rates of return.

One way of describing the causes that lie behind such phenomena is as follows. The accounting rate of return in exponential growth depends on the time shapes of benefits that flow from investments and on the rate of growth of the firm. It is true, given the rate of growth and the general appearance of the time shape of benefits, but not its level, that the accounting rate of return and the economic rate of return will be positively associated, but that association is not strong enough to allow one to ignore the obviously realistic possibility that growth rates and time shapes differ over firms. Particularly in the case of time shapes, so little is actually known that one assumes away differences at one's peril (and, I may add, if one knows the time shape of benefits including its level, one knows the economic rate of return without further ado).

I should have thought that, given such results, the burden of proof would be on those who wish to continue in the belief that accounting rates of return can be used as valid indices of economic profit rates. Surely, more is called for than the simple assertion that our examples are unrepresentative and that it is well known that accounting rates of return measure profits. That is particularly so because, on the contrary, it is—or ought to be—well known that this is not the case. Many of our results were not new, and, as we pointed out, many of the same points had been made by others.[2]

II. End-of-Year vs. Beginning-of-Year Rates

I now turn to the specific comments published above, concentrating on that of William F. Long and David J. Ravenscraft (L-R). As do L-R, I begin with what may at first seem a minor matter, the treatment of end-of-year vs. beginning-of-year assets.

Long and Ravenscraft assert that McGowan and I incorrectly calculated accounting rates of return on end-of-year assets. They claim that this must be so because our end-of-year-assets rates of return are lower than or equal to our beginning-of-year-assets rates of return and state: "If there is depreci-

[2] I wish to take this occasion to rectify an inexcusable scholarly oversight on our part. While we cited a number of predecessor writings, we failed to cite that of G. C. Harcourt (1965); this was particularly unfortunate because of all the literature, Harcourt's valuable article is perhaps the one most closely related to our own work. In this connection, it is perhaps useful to point out that J. A. Kay's criticism (1976) of Harcourt is quite misleading. For example, Kay states (result (*i*), p. 449) that if the accounting rate of return on a particular project is constant over the life of that project, then that accounting rate of return equals the internal rate of return. He fails to note that such constancy can only occur if Hotelling valuation ("economic depreciation") is used (see our article, Theorem 1, p. 91 and also p. 93, fn. 24). More important, Kay's calculation of the economic rate of return for the firm as a whole from a time-series of accounting rates and a terminal valuation either requires that the firm be wound up, in which case, all that is involved is a direct knowledge of the time shape of benefits or else requires that the terminal valuation used be Hotelling valuation which requires knowledge of the economic rate of return. See also F. K. Wright (1978).

ation, and if the same accounting profit value is divided by the two asset values, the end-of-year accounting rate of return must be larger than the beginning-of-year accounting rate of return" (p. 494). This is flatly false. So long as the firm has positive net investment, end-of-year assets will exceed beginning-of-year assets. Nonnegative net investment occurs in our computer-generated examples for every case of exponential growth, and positive net investment for positive rates of growth. It is not hard to see that, in exponential growth, the accounting rate of return on beginning-of-year assets will exceed that on end-of-year assets by a factor of one plus the growth rate, and this is true of the results given in our paper. It is not true of the supposedly "corrected" results given by L-R.

This part of L-R's treatment of the beginning- vs. end-of-year assets question may be only indicative of their keen sensitivity even in trivial respects to a paper that criticizes the accepted way of doing things in this area, but the remainder of that treatment is symptomatic of a more fundamental misunderstanding. They claim that the fact that accounting rates of return on end-of-year assets perform even less well in our examples than accounting rates of return on beginning-of-year assets is an artifact stemming from the fact that we defined the economic rate of return in beginning-of-year terms. They observe:

> The continuous time results derived in F-M's Appendix hold in discrete time for accounting profit rates defined with beginning-of-year assets as the denominator, if the growth rate and internal rate of return are defined in beginning-of-year terms. However, it [*sic*] also holds for accounting profit rates defined with end-of-year assets as the denominator, provided the growth rate and internal rate of return are defined in end-of-year terms.
>
> [p. 494–95]

They also point out that suitable redefinitions will make our results hold for accounting rates of return defined on any given convex combination of beginning- and end-of-year assets.

This is certainly true[3] and just as certainly completely irrelevant. What L-R fail to realize is that the issue is not whether one can redefine the magnitudes of economic theory (here the economic rate of return) so that certain theorems (our results) will be true of their relations to the magnitudes used in practice (here accounting rates of return on end-of-year assets). Rather the issue is whether the magnitudes used in practice bear any useful relationship to the magnitudes of economic theory. The fact that one can invent something called the "end-of-year internal rate of return" and show that its relations to the accounting rate of return defined on end-of-year assets are such as to obey the theorems in our Appendix is not a reason for using the accounting rate of return defined on end-of-year assets.

This is an important point and it is worth going into in more detail. The economic rate of return is the magnitude which gives the signal for entry or exit of resources. That is because firms can compare the economic rate of return offered by a particular investment opportunity with the cost of capital. At least in those cases in which rate of return calculations are appropriate at all, such comparisons will yield the correct actions for long-run profit maximization.[4]

Now consider the internal rate of return which L-R propose to define on end-of-year assets (given in fn. 3 above). That rate is undefined for assets such as one-year bonds whose payoff comes entirely in one year. That means that firms deciding between one-year bonds and other assets cannot use the L-R-defined rate of return to make that decision. More generally, that rate of return cannot be used to decide among different

[3] L-R's definition of the appropriate internal rate of return to use on end-of-year assets is given in their fn. 2 and elaborated in their working paper (1983).

[4] As is well known, there are occasions on which rules based on rate of return comparisons are not equivalent to present value maximization which always yields the correct rule, but in such circumstances the literature which uses accounting rates of return to make inferences about profits cannot possibly be correct. Contrary to what some of my correspondents appear to believe, McGowan and I were not recommending internal rate of return calculations as a substitute for present value maximization.

investments that have significant positive payoffs in the first year or to compare such investments with others; indeed, if one examines L-R's general formula, one finds that any two-period investment with positive cash flow in both years and the cash flow in the first year greater than the investment cost has an L-R rate of return greater than one and is to be preferred to any two-period investment with first-year cash flow below the investment cost *no matter what the sizes of the second year cash flows are*. Similar anomalies arise for longer cash flow profiles. In the light of this, what difference does it make whether or not such an internal rate of return relates to the accounting rate of return on end-of-year assets in the way given for continuous time rates in our Appendix?

I have discussed this in some detail because it is symptomatic of the reluctance on the part of those involved in the use of accounting rates of return to take a serious look at what it is that such rates of return are supposed to be measuring. Long and Ravenscraft are so anxious to rescue accounting rates of return on end-of-year assets that they make an elementary error (inconsistently treating net investment) and have apparently forced their calculations to show the impossible. Worse, they show that they can define magnitudes (their redefined internal rates of return) which relate to the particular version of accounting rates of return they wish to defend in the same (not very useful) way that the usual internal rate of return relates to the accounting rate of return on beginning-of-year assets; but they fail to notice that the magnitudes they define no longer have the property that the usual internal rate of return does of being something in which the analysis is interested.[5]

III. The Burden of Proof and the Role of Examples

The same tendency to defend rather than consider runs through much of the rest of L-R's criticisms. As already remarked, given our results, one would think that the burden of proof was on those who wish to go on using accounting rates of return as measures of economic profit. Since our examples merely illustrate some general theorems, that burden cannot be sustained by arguing that our examples are unrealistic; one must show rather than presume that the problems do not arise for real firms.

In this connection, it is instructive to consider L-R's valid point that since our examples considered the accounting rate of return on total assets rather than on equity (the firms in our examples had no debt), our use of 15 percent as the value of the economic rate of return was incorrect. If we were going to choose a rate representative of accounting rates of return, we should have used the average accounting rate of return on total assets (7.8 as opposed to 15 percent in 1978). Quite so. But consider what they say next:

> If an economic rate of return of 7.8 percent is used instead of 15 percent, and the set of growth rates is centered on 7.8 percent, the maximum deviation from the economic rate of return on beginning-of-year assets is 3.9 vs. 10.9 percentage points in F-M's Table 2 or 50 vs. 73 percent of the economic rate of return. [p. 495]

Apparently, a 50 percent error is a small one. Moreover, while it is undoubtedly true that the spread of accounting rates of return in such an exercise would be reduced *for these particular examples*,[6] it must be remembered that the examples in our Table 3 are all variations on a particular theme suggested by an earlier critic who thought that earlier examples generating a still wider range were chosen to be unrepresentative (see our article, p. 85, fn. 14, and our book with Joen Greenwood, p. 242). Other, more widely varying examples would yield an even wider spread, so that no comfort can be taken from the prospect that accounting rates of return may be "only" 50 percent in error.

[5] There are, of course, problems involved in discrete time models as to the handling of cash flows that come in and investments that are made in the course of a year, but they are not the ones being discussed here.

[6] Whether the particular figures given by L-R can be trusted in light of their results on end-of-year accounting rates is a different matter.

Similar remarks apply to L-R's contention that our examples are misleading because most of them involve accelerated depreciation. The spread of accounting rates of return for a given economic rate of return is quite large enough in the straight-line depreciation cases given in our Table 2 to show that accounting rates of return cannot be used as measures of economic rates of return. Furthermore, that spread would be wider still for different examples.[7]

To sum up: our examples only illustrate the general theorems involved. Only those who are determined to believe without any serious theoretical basis that accounting rates of return measure economic rates of return can find L-R's arguments persuasive.

IV. Correlation Between Accounting and Economic Rates of Return

Plainly, what is needed is an affirmative showing that accounting rates of return in fact measure what they are supposed to, and L-R do indeed make some gestures in that direction. They suggest that accounting rates of return and economic rates of return are likely to be correlated; they make some "indirect" tests of whether conclusions based on accounting rates of return are likely to be in error; and they claim that accounting rates of return meet the market test of being widely used.

It would be surprising if accounting rates of return were not correlated at all with economic rates of return. As pointed out above, given the growth rate of the firm and, especially, given the general appearance of the time shape of benefits accruing from a typical investment, then (at least in exponential growth) the accounting rate of return will be positively associated with the economic rate of return because both will be positively associated with the level of the benefit stream. It is this fact which allows management of particular companies with reasonably constant time shapes and growth rates to use accounting rates of return as "hurdle rates" as discussed by Michael van Breda in his comment. It would be surprising to find that such positive association fails to persist when time shapes and growth rates vary over firms; one expects variables with a positive coefficient in a multiple regression to show some positive zero-order correlation with the dependent variable.

How strong is that correlation in fact? That is very hard to know, precisely because we do not know the economic rates of return of most firms (and if we did, any information provided by the positive correlation in question would be superfluous). Long and Ravenscraft cite Thomas Stauffer to the effect that the correlation is .79 over nine selected industries and speculate that this understates the correlation for all industries. Unfortunately, the basis for this is quite weak—and inevitably so. Stauffer's estimation of internal rates of return for the nine industries studied rested (as was inescapable) on specific assumptions about the time shapes of benefits which he thought characterized those industries. Those assumptions were plausible, but they could not be precise calculations. Since we know that the relation between the accounting and economic rate of return is very sensitive even to relatively small variations in the assumed time shape (see our Table 3, p. 87), the correlation which L-R derive from Stauffer's results can only be taken as somewhat suggestive. No general conclusion can be drawn.

More important than the exact size of the correlation, however, is the fact that it plainly is not very close to one. Given that, the issue is not how high the correlation is but whether differences between accounting and economic rates of return are related to variables used in statistical studies. If they are, then studies which use the accounting rate of return as a proxy dependent variable for the economic rate of return are likely to give misleading or unreliable results.

This point is well made in a pair of studies by Gerald Salamon (1983) and by Salamon and Mark Moriarty (1983). They observe that McGowan's and my results show that the choice of a depreciation method affects the relation between the economic rate of return and the accounting rate of return;

[7] And, I may add, large firms—including those that may be suspected of having monopoly power—tend to take accelerated depreciation.

since there is evidence that the choice of depreciation method is related to firm size, conclusions as to the relations between firm size and profitability which stem from the use of the accounting rate of return must be viewed with caution. Salamon estimates economic rates of return for a selection of firms (assuming a particular parametric family of time shapes) and shows that conclusions on the size-profitability relation are indeed affected. He and Moriarty reach a similar result as to the conclusions of the literature on the relations of advertising and profitability. While these results depend on the assumptions as to time shape, it would bę perilous indeed to conclude that they are not general. Only very detailed investigation can rescue the accounting rate of return and the studies based on it.

V. The L-R "Indirect" Evidence

Long and Ravenscraft do not offer such detailed investigation; instead they offer indirect evidence. Some of that evidence is indirect indeed. The fact that distortions in line-of-business profits coming from common cost allocations or nonmarket transfer prices may leave line-of-business profits correlated with profits measured differently has no bearing whatever on the problems at issue here. More important, evidence relating to studies using profit rates on *sales* is similarly irrelevant.

I cannot understand how even the most casual reader of our paper could think that McGowan and I were discussing any accounting rate of return other than that on assets or on stockholders' equity. In particular, our results plainly have nothing to do with the usefulness or lack thereof of the accounting rate of return on sales. Yet L-R treat this as though it affected the relevance of our results, and Stephen Martin in his comment uses our paper as an occasion to write about the relations between the accounting rate of return on sales and the Lerner measure of monopoly power as though such relations bore on what we did discuss. I shall comment on this matter below; for the present, however, I wish to stay with the question of the relevance of the L-R discussion of results involving the rate of return on sales to the applicability of our results.

Long and Ravenscraft calculate regression statistics for a structure-profits regression using profits/sales and gross profits/gross assets as alternative dependent variables. They find similar results in both regressions.[8] They then reestimate using profits before depreciation in the numerators of the dependent variables and attempt to interpret the results as to the likely effect of the distortions introduced by different depreciation methods on the relations between the accounting rate of return and the economic rate of return. They find in the case of profits/sales as the dependent variable that five out of twenty-three coefficients are affected (including the coefficient of minimum efficient scale as Salamon's results would suggest). Changes in the case of the gross profits/gross asset regression were even less significant. L-R appear to draw some comfort from these results.

It is unnecessary to consider whether five out of twenty-three is large or small. These results are meaningless. Consider first the results using profits/sales as the dependent variable. Adding depreciation back into the numerator of the dependent variable produces a variable with no content in terms of economic analysis. Depreciation, however difficult its computation, represents a real economic cost. Profits before depreciation no more relate to the Lerner measure than do profits before subtraction of labor costs. To the extent, therefore, that the L-R results do not change when gross profits are used, one must question the meaning of their results when net profits are used in the dependent variable. As it is, all that can be concluded from these results is that depreciation/sales in the L-R sample is related in particular ways to at least five of the variables they use.

The case as regards the regressions using gross profits divided by assets is similar. Such a "depreciation-adjusted" accounting

[8] It is typical of L-R's underlying attitude toward the issues that arise when the dependent variable used may not measure what it is supposed to measure that they should state: "the strongest statistical results arose in the profits/sales regression, which lends support to the choice of profits/sales over profits/assets as the dependent variable in such regressions" (p. 495).

rate of return is no more likely to be closely related to the economic rate of return than is the unadjusted accounting rate. Indeed, since depreciation represents real costs, such an adjusted rate of return makes even less analytical sense than do the usual measures.[9] To find that such an adjustment leaves results unaffected is to cast grave doubt on the results if it says anything at all.

The results reported by L-R using accounting rates of return on end-of-year or middle-of-year instead of beginning-of-year assets are likewise irrelevant. Since none of these measures correctly reflects the economic rate of return, the most that can be concluded from the fact that results using each are similar is that the differences among them are less important than the differences between the economic rate of return on the one hand and such measures as a group on the other. Since accounting rates of return on beginning-of-year and on end-of-year assets are likely to be highly correlated,[10] any other result would be surprising, whatever the facts as to the economic rates of return of the firms involved.

VI. The "Market Test"

I now turn to the question of the market test of acceptance of accounting rates of return. Here again, L-R are quite confused. McGowan and I attacked the use of accounting *rates of return*—L-R behave as though we had attacked the use of data on accounting *profits*, and that is simply beside the point. Accounting profit data have their problems as measures of economic profits, but those problems are relatively well understood and were certainly not the subject of our paper. It is unnecessary further to comment on L-R's discussion leading up to their somewhat hysterical comment that "the implication of F-M's work, if correct, is that most of applied economics is misguided" (p. 495).

The only market test that is of any relevance, then, concerns the direct use of accounting rates of return themselves. That use is explained along the lines given above as to the reasons for positive correlation between accounting and economic rates. In the absence of better information, firms (and investment houses) may use accounting rates of return as rules of thumb—"hurdle rates" as van Breda puts it. Careful comparisons for a given firm over time or among firms in the same industry may yield some rough information because growth rates and benefit profiles may be roughly the same (although even this is open to question). But, if accounting rates of return really measured economic rates of return, there would be no need for potential investors or financial analysts to use anything else, and they plainly do, examining, among other things, rates of growth and, especially, prospective cash flows. In any event, such use of accounting rates of return by the lay public cannot substitute for analysis of their properties by the economist.

VII. The Lerner Measure and the Rate of Return on Sales

I now leave the discussion of Long and Ravenscraft and comment briefly on issues which arise in the rather more sensible comments of the other authors involved in the above discussion. The first such subject, as has already been indicated, is that of the use of profits divided by sales as a measure of monopoly profits. This is the subject—indeed the only subject—of Stephen Martin's comment.

I have already indicated that this topic seems to me irrelevant as a criticism of McGowan and me. Our paper was plainly about the use of profits divided by assets or by equity. To reply that profits divided by sales may be a useful measure is not to reply at all. This is therefore not the appropriate place to discuss the failings of the return-on-

[9] L-R have at least avoided the pitfall of adding depreciation back to the numerator of the accounting rate of return while leaving depreciated assets in the denominator. Such an adjustment is not unknown to those attempting to make inferences from accounting rates of return; it was made by Alan McAdams in his testimony for the government in the IBM case. Since IBM took relatively faster depreciation than did most of the computer firms with which it was being compared, the results were predictable as an arithmetic artifact. See our book with Greenwood, pp. 236, 257.

[10] Indeed, as pointed out above, in exponential growth the two rates of return will be proportional for firms with the same growth rate.

sales-as-a-measure-of-monopoly-power literature. Since the topic has been brought up, however, I shall add a few remarks.

1) Nobody supposes that profits divided by sales are an interesting profit rate measure because they tell one something about the desirability of moving resources in or out of a given area. If profits divided by sales are interesting, it is because, under constant returns to scale, the rate of return on sales is related to the Lerner measure of monopoly power, not because that rate of return is truly a profit rate in the sense that economic theory uses that term. I stand by our statement that "the economic rate of return is the only correct measure of the profit rate for purposes of economic analysis" (p. 82).

2) The use of profits divided by sales as reflective of the Lerner measure is something which needs to be approached with considerable caution. It seems to me quite dangerous to assume constant returns when dealing with analyses involving firm size and concentration.

3) Even the Lerner measure itself seems to me to be of limited usefulness. There is a clear sense in which, for a given firm and (possibly) for a given industry, the higher the Lerner measure the greater the monopoly power, but it is not so clear in what sense this is true when different firms or industries are to be compared as in statistical studies. There is no natural metric here and it is not clear how high is high, or even whether higher is necessarily more powerful.

4) This can be related to the first point made above by observing that the Lerner measure is not directly related to the economic rate of return. An industry with a high Lerner measure and a low economic rate of return does not strike me as ripe for antitrust action; an industry with high economic rate of return which is unaccounted for by any reason other than the possible presence of monopoly[11] does so strike me, even if it has a low Lerner measure. In this sense I would include even the rate of return on sales (which we certainly did not have in mind) in the statement by McGowan and me that "Accounting rates of return are useful only insofar as they yield information as to economic rates of return" (p. 82). Whether the Lerner measure has that property is an open question.

VIII. Concluding Remarks

I can, I think, do no better than to conclude with a brief comment on Ira Horowitz. His paper is at least amusing, and there is no point in being overly solemn about it. Nevertheless, to the extent that it is to be taken seriously, one searches in vain for any argument that McGowan and I were mistaken. Rather, the underlying assumption again appears to be that everyone knows that high accounting rates of return are indicative of monopoly power and that therefore this must be true.

I thus read Horowitz as saying that McGowan and I have made it harder for economists to give loose but pontificating testimony for which there is no solid analytic foundation. Particularly considering the origins of our paper in *U.S. v. IBM*,[12] I am not going to lose any sleep over that. Certainly, John McGowan would have considered it the highest possible praise.

[11] This is by no means an easy thing to discover. There are many other roles for and sources of profits besides monopoly. See our book with Greenwood, ch. 7.

[12] See our book with Greenwood, ch. 7.

REFERENCES

Fisher, Franklin M. and McGowan, John J., "On the Misuse of Accounting Rates of Return to Infer Monopoly Profits," *American Economic Review*, March 1983, *73*, 82–97.

______, ______, and Greenwood, Joen E., *Folded, Spindled, and Mutilated: Economic Analysis and U.S. v. IBM*, Cambridge: MIT Press, 1983.

Harcourt, G. C., "The Accountant in a Golden Age," *Oxford Economic Papers*, March 1965, *17*, 66–80.

Horowitz, Ira, "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, *74*, 492–93.

Hotelling, Harold, "A General Mathematical Theory of Depreciation," *Journal of the American Statistical Association*, September 1925, *20*, 340–53.

Kay, J. A., "Accountants, Too, Could Be Happy in a Golden Age: The Accountant's Rate of Profit and the Internal Rate of Return," *Oxford Economic Papers*, November 1976, *28*, 447–60.

Long, William F. and Ravenscraft, David J., "The Usefulness of Accounting Profit Data: A Comment on Fisher and McGowan," Working Paper No. 94, Federal Trade Commission, June 1983.

______ **and** ______, "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, *74*, 494–500.

Martin, Stephen, "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, *74*, 501–06.

Salamon, Gerald L., "Accounting Rate of Return, Measurement Error, and Tests of Economic Hypotheses: The Case of Firm Size," mimeo., June 1983.

______ **and Moriarty, Mark M.**, "Alternative Profitability Measures and Tests of Economic Hypotheses: An Application to the Advertising-Profitability Issue," mimeo., May 1983.

Stauffer, Thomas R., *The Measurement of Corporate Rates of Return and the Marginal Efficiency of Capital*, New York: Garland, 1980.

Van Breda, Michael F., "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, *74*, 507–08.

Wright, F. K., "Accounting Rate of Profit and Internal Rate of Return," *Oxford Economic Papers*, November 1978, *30*, 464–68.

Market Opportunities, Genetic Endowments, and Intrafamily Resource Distribution: Comment

By NANCY R. FOLBRE*

In their recently published article in this *Review*, Mark Rosenzweig and T. Paul Schultz (1982) provide an imaginative and empirically compelling account of resource allocation in Indian households. They argue that differences in relative female-male infant mortality levels are attributable to differences in intrafamily resource distribution motivated by the larger potential contribution of female children to family income in areas with higher levels of female labor force participation. In other words, families constrained by poverty pursue an "orderly optimizing process," giving female infants less food and medical care than male infants unless they anticipate that daughters will be able to make a contribution to family income comparable to that of sons. This argument is likely to offend those who reject the premise that family behavior is influenced by economic concerns. But, by providing an explanation of inequalities within the household that often go unexplained, Rosenzweig and Schultz have substantially enhanced the credibility of the "new home economics."

The notion that unequal treatment of female children is the result of an orderly optimizing process on the part of the family is appealing precisely because no other economic explanations have been offered. Unfortunately, the optimizing process which the authors describe is not as orderly as it seems. Strict adherence to the neoclassical economic assumption that individuals are optimizers reveals serious inconsistencies in the new home economics assumption that family behavior is shaped by an exogenously given joint utility function. These inconsistencies center around the possibility that the joint utility function is not exogenous to the family, but is instead shaped within it by the economic power that individual optimizers have to impose their own tastes and preferences on other household members.

There is no a priori reason to dismiss the possibility that unequal treatment of females is embodied in a joint utility function that places a lower value on female utility than on male utility. Indeed, one leading proponent of the new home economics writes that the presumption of an unequal joint utility function fits quite well in the "male dominated families of much of the low income world" (R. E. Evenson, 1976, p. 88). Rosenzweig and Schultz never address this issue, despite their reference to "exogenous cultural factors (for example, religion and caste)" (p. 807), which amounts to a form of discrimination against women in the labor market. In fact, labor market discrimination against women is central to their argument for in its absence we might assume, following Reuben Gronau (1973), that females do not join the labor force because their marginal product in household production is greater than the market wage. Thus their contribution to the household would not necessarily be lower than that of males, although it would take the form of goods and services, rather than market income.

If discrimination exists in the labor market, despite the efficiency losses it entails there, is it necessarily absent from the household? Is the impulse to altruism within the household greater than the incentive to efficiency within the capitalist firm? Do employers (primarily male) have one set of tastes and preferences in the labor market and another in the home? Or are female infants that receive fewer calories and less protein than male infants being discriminated against?

The new home economics approach can accommodate the possibility that a "taste for discrimination" is embodied within the household's joint utility function. As long as the postulated joint utility function is exogenous and constant, differences in household

*Department of Economics, Graduate Faculty, New School for Social Research, New York, New York 10003.

behavior can be explained as an optimizing response to differences in market opportunities. But, even if individual tastes and preferences are exogenous and constant over time, the way in which these individuals' tastes and preferences are aggregated within the household's utility function may change.

There are a number of reasons why the joint utility function may change as a result of changes in prices and incomes. If individual women are optimizers, they will join a household that places a lower value on their individual utility than on the utility of men as long as there are no more attractive opportunities. But, as opportunities for women to earn income outside the household become available, their bargaining power may increase and enable them to modify the household's joint utility function. Marjorie McElroy and Mary Jean Horney (1981) as well as Marilyn Manser and Murray Brown (1980) have formalized such an approach through the use of Nash-bargaining models.

The infants whose life chances are determined by intrafamily resource distribution presumably do not exert any influence on the joint utility function (if they did, one wonders how female infants would weigh the utility of additional income for the household relative to the utility of their own survival). But what happens as children age? Do parents retain their ability to impose their own tastes and preferences on their children as they become adults? Or is such an imposition partly contingent upon the economic power that parents have over children, the power to impose altruism on "rotten kids" (Gary Becker, 1981)?

The economic benefits of children are determined not only by their earnings or their product, but also by the joint utility function that determines their propensity to share it with their parents. If one assumes, as Rosenzweig and Schultz implicitly do, that this propensity remains unchanged, one may assume that increases in market income lead to increases in filial contributions. But historical evidence indicates that increases in market opportunities for young men ultimately lead to a reduction in the flow of income to the older generation (John Caldwell, 1982). Increasing market opportunities for young women may initially have the opposite effect since they often lead to an increase in age at marriage and prolongation of coresidence and income sharing with their parents (Janet Salaff, 1981).

Rosenzweig and Schultz muster no evidence to show that the joint utility function is exogenously given and their empirical results are consistent with an explanation that differs substantially from their own. Mothers in Indian households may benefit relatively more from the survival of female children than fathers do, since daughters commonly assist their mother in household tasks. Greater market opportunities for women as wives may augment their bargaining power, allowing them to modify the joint utility function and to increase the flow of resources to female children. Furthermore, increased market opportunities outside the household may increase young men's propensity to exit from or modify the households' joint utility function and thereby lower their desirability as an "investment" relative to young women. Rosenzweig and Schultz's empirical analysis would not capture this effect because it does not distinguish between market opportunities that are independent of the parental household and those that involve use of parental land or capital.

Many noneconomists have expressed skepticism regarding the basic assumptions of the neoclassical approach to household decision making. Such skepticism may or may not be warranted. I have argued above that the argument presented in Rosenzweig and Schultz's article is inconsistent on its own terms, for it suggests that households respond to economic factors to seek an optimal allocation of resources, but that individuals do not respond to economic factors to seek an optimal joint utility function. Few scientific endeavors can proceed without reliance on somewhat heroic assumptions. But such assumptions must be applied consistently, especially where their modification leads to an entirely different interpretation of the same empirical results.

Both interpretations of the Indian data reach a similar conclusion—that the expansion of market opportunities may mitigate inequality in the household. But the paths to this conclusion diverge from the very outset. In Rosenzweig and Schultz's view, differences in intrafamily resource allocation reflect responses to economic factors which are optimal from the point of view of the household as a whole. The alternative view is that such responses may be more optimal for some family members than for others. Again, in Rosenzweig and Schultz's view, the expansion of market opportunities automatically leads to declining inequality within the household. The alternative view is that market opportunities have this effect only if they empower disadvantaged household members to change the joint utility function. Rosenzweig and Schultz's analysis mirrors the larger neoclassical vision of a world in which no significant conflict of interest exists. However seductive such a world may be, it bears little resemblance to the one inhabited by Indian families or by U.S. economists.

REFERENCES

Becker, Gary, *A Treatise on the Family*, Cambridge: Harvard University Press, 1981.

Caldwell, John C., *The Theory of Fertility Decline*, New York: Academic Press, 1982.

Evenson, R. E., "On the New Household Economics," *Journal of Agricultural Economics and Development*, January 1976, *6*, 87–103.

Gronau, Reuben, "The Intrafamily Allocation of Time: The Value of Housewives' Time," *American Economic Review*, September 1973, *63*, 634–51.

McElroy, Marjorie and Horney, Mary Jean, "Nash-Bargained Household Decisions: Toward a Generalization of the Theory of Demand," *International Economic Review*, June 1981, *22*, 333–49.

Manser, Marilyn and Brown, Murray, "Marriage and Household Decision Making: A Bargaining Analysis," *International Economic Review*, February 1980, *21*, 31–43.

Rosenzweig, Mark and Schultz, T. Paul, "Market Opportunities, Genetic Endowments, and Intrafamily Resource Distribution," *American Economic Review*, September 1982, *72*, 803–15.

Salaff, Janet, *Working Daughters of Hong Kong: Filial Piety or Power in the Family?*, New York: Cambridge University Press, 1981.

Market Opportunities, Genetic Endowments, and Intrafamily Resource Distribution: Reply

By MARK R. ROSENZWEIG AND T. PAUL SCHULTZ*

In her comment, Nancy Folbre accepts our empirical finding that more household resources in India are allocated to female infants where labor market opportunities for women are greater. But she provides an alternative explanation. We employ a model with a joint family utility function that is independent of market forces, and explain the empirical regularity as a consequence of the model's prediction that parents allocate resources to those uses which produce the highest returns. Folbre rejects a priori the exogeneity of the joint household utility function and proposes an alternative rationale for our finding consisting of *two* hypotheses: 1) where women can provide more resources to the family or have relatively greater market earnings opportunities, their distinct preferences have greater weight in household decisions; and 2) mothers prefer to allocate more resources to girls compared to what fathers would like to allocate.

Whatever the inherent merits of modeling household resource allocations as the outcome of a bargaining process, to use this general approach to explain our empirical results evidently requires Folbre to impose more assumptions about differences between men and women than is required in the joint family model. In our model men and women differ importantly only in the market rewards for their services (which may arise from differences in inherent productivity or from market discrimination or both), and possibly in their capacities to utilize nutrients; in Folbre's framework, not only are market rewards different for men and women, but there are persistent and important differences in the preferences of men and women for goods and services, which evidently even assortative mating does not eliminate between marital partners. And, for reasons not made clear, women prefer girls more than men do, independent of investment returns. Folbre merely substitutes "exogenous and constant" and particular taste differences between the sexes for our agnostic "exogenous" household utility function (in which women may also prefer girls to boys).

Given the lack of any direct evidence presented that documents the tastes differences that form the core of Folbre's alternative explanation, the issue of model choice is whether the proposed increase in complexity can be justified by the generation of additional predictions or predictions different from those obtained from our model. Folbre's remarks would be more useful if they could guide further empirical investigation or data collection.

One well-known test of the hypothesized joint family utility function, for example, is to examine if the ownership or control of resources (such as financial assets) by husband and wife is significantly related to household resource allocations. Marjorie McElroy and Mary Jean Horney (1981), indeed, formally develop an alternative household bargaining model within which they can nest the joint utility demand system. Folbre's framework appears to imply that, for given labor market opportunities for men and women, in households where women own a larger share of resources, perhaps through inheritance, allocations to girls in relation to boys would be larger. Thus, a proper test of this hypothesis would appear feasible if data were collected on 1) asset ownership, or perhaps more importantly, the control of resources by household members, 2) direct observations on intrafamily household resource allocations between boys and girls, or indi-

*Departments of Economics, University of Minnesota, 1035 Management and Economics, Minneapolis, MN 55455, and Yale University, New Haven, CT 06520, respectively.

rect indications of these allocations, as we exploited sex-specific child mortality in India, and 3) measures of exogenous market forces outside of the household's control that affect the returns to intrahousehold allocations.

Logically, of course, one can reject empirically (or out of hand) the joint utility maximization model and still retain the prediction that men or women allocate resources across activities or individuals according to their relative market-determined returns. Folbre's second hypothesis is that, presumably for given market opportunities, mothers are more willing than fathers to allocate their resources to girls. It is not clear, however, whether Folbre believes that the distinct "taste" for allocating resources to girls on the part of mothers is an exogenous and universal biological endowment, or whether this preference depends endogenously on attributes of the household technology. Mothers in Indian households are said to find daughters better substitutes for themselves in the "tasks" they perform. To confirm this line of reasoning, one would need to obtain evidence on the substitution elasticities between mothers,' daughters,' and sons' time inputs into household production. But, of course, changes in adult sex-specific market opportunities would also generate intrafamily substitution between mothers and daughters in household tasks within a model of joint utility maximization.

If the joint family utility framework is to be replaced by a less parsimonious model of intrafamily resource allocation, the increase in complexity should be explicitly demonstrated to have empirically distinguishable predictions. Developing a research agenda that will facilitate modeling conflict of interest situations between men and women, and between generations, that are resolved within the family is certainly an important challenge that has attracted the attention of too few economists. But progress in this direction will require sharply focused conceptualizations and empirical research, not just the expression of dissatisfaction with the tractable simplicity of the joint family function, or a distaste for its implications.

REFERENCES

Folbre, Nancy R., "Market Opportunities, Genetic Endowments, and Intrafamily Resource Distribution," *American Economic Review*, June 1984, *74*, 518–20.

McElroy, Marjorie and Horney, Mary Jean, "Nash-Bargained Household Decisions: Toward a Generalization of the Theory of Demand," *International Economic Review*, June 1981, *22*, 333–49.

The Political Economy of Political Philosophy: Comment

By STEVEN A. COBB AND ROBERT P. HAGEMANN*

In a paper recently published in this *Review*, James Bennett and Thomas DiLorenzo (hereafter, B-D) estimate a model designed to "test the hypothesis that conservatives, *ceteris paribus*, do in fact return a larger proportion of their staff allocations unspent than liberals" (1982, p. 1160). While they make no claim to be modeling efficient management behavior, they are sufficiently confident in their ability to control for what might be variously labeled as state effects, job effects, and political effects, to conclude that "senators who are conservative in making collective decisions are also fiscally conservative in spending public funds under their direct control" (p. 1160). Their regressions indicate that several such factors are significantly related to the percentage of staff funds returned by each senator and hence may influence the estimated relationship between measures of political and fiscal conservatism. However, B-D do not effectively control for these variables. In this comment we suggest a natural alternative specification of their model which automatically controls for all state effects. If the B-D model were correct, our specification would produce parameter estimates statistically indistinguishable from those obtained by B-D. However, we find that this is not the case and, in particular, under our specification, the relationship between political and fiscal conservatism disappears.

To measure senatorial frugality, B-D suggest using the treatment of clerk-hire budgets by the senators of the 95th Congress. These budgets, allocated to staff each senator's offices on the basis of state population, cannot be carried over from one fiscal year to the next. Any money not spent must therefore be returned to the Treasury. Some politically conservative senators have publicized the fact that they return a relatively large percentage of their clerk-hire allowance. While making no judgment on the prudence of such behavior, B-D correctly note that other factors may influence the percentage of the budget which a senator may return. In particular, they suggest variables which may influence the cost of staffing an office, including 1) the number of committees on which the senator serves, 2) the number of chairmanships held by the senator; 3) the number of years that the senator has held office; 4) whether or not the senator is up for reelection; and 5) a large number of variables describing the state which the senator represents (for example, area of state, per capita income, number of families, etc.) In spite of the fact that B-D tried unsuccessfully to include a number of additional state-related effects, it appears that they failed to take into account other such variables which may, in fact, confound the estimated relationship. A noteworthy example, which might be termed an allocation effect, may be seen by examining the clerk-hire allocation schedule (see B-D, Table 1, p. 1153). This schedule reveals that the state per capita senatorial staff allowance falls as state population increases. Although the costs of staffing an office undoubtedly depend, *ceteris paribus*, upon the size of the state population served,[1] it seems likely that allocation errors which vary with state population may exist. Depending upon the relationship between such errors and state population on the one hand, and political conservatism on the other, it is possible to obtain a spurious relationship between fiscal and political conservatism when state effects are not fully taken into account.

*Economists, Bureau of Labor Statistics, 600 E Street, NW, Washington, D.C. 20212, and Department of the Treasury, 15th & Pennsylvania Avenue, NW, Washington, D.C. 20220, respectively. We thank Robert Gillingham, John Greenlees, and Edward Murphy for their suggestions. The views expressed are our own and do not represent those of the BLS, the Treasury Department, or other members of their staffs.

[1] For example, a linear regression of staff expenditures upon state population levels "explains" 68 percent of the variation in the dependent variable.

In their specification, B-D estimate a linear regression for the 96 full-term senators in the 95th Congress and find statistically significant coefficients for several control variables including any of four highly correlated measures of political conservatism.[2] They therefore conclude that politically conservative senators tend to return a higher percentage of their clerk-hire budget. However, given their emphasis on interpretation of the political conservatism variable, we suggest an alternative specification that, although generating no state-effect coefficient estimates, has the desirable property of providing an estimate of the conservatism coefficient while automatically controlling for *all* state and allocation effects. Our specification requires less data and, under the maintained hypothesis of the B-D model, provides unbiased and consistent estimates for all included variables. Our alternative is possible because their data set includes two senators from each state and consequently a regression estimated on differences *between* state senators will automatically purge the regression of all state effects. Providing that sufficient variation remains in the differenced dependent and independent variables, the resulting parameter estimates, while less efficient, will not differ in expectation from those in the B-D regression whenever the B-D regression is correct.

Using all the nonstate variables included in the B-D regression, we estimate a linear regression for the 92 senators of the 95th Congress representing states in which both senators served for the complete period. Because each observation is on the within-state differences for these senators, we actually have 46 independent observations. However, we only need 5 independent regressors because the state-specific variables drop out.[3] In the results presented below, the nonstate variables, including the differences in the paired senators' *ADA* ratings, explain only about 11 percent of the variation in the dependent variable.

$$\begin{aligned} \mathit{PERCENT\ RETURNED} = {} & .189 \\ & + \underset{(0.065)}{.056}\,\mathit{ADA} + \underset{(1.783)}{.804}\,\mathit{COM} + \underset{(4.273)}{8.611}\,\mathit{CHAIR} \\ & + \underset{(2.669)}{1.972}\,\mathit{REELECT} - \underset{(0.217)}{.174}\,\mathit{TEN} \end{aligned}$$

$$R^2 = .113;\ F(5,40) = 1.019; \text{obs} = 46.$$

In this equation, *PERCENT RETURNED* is the percent of clerk-hire budget returned based on 1976 state population, *ADA* is 100-*ADA* rating for the second session of the 95th Congress, *COM* is number of committees served, *CHAIR* is number of committees chaired, *REELECT* is a dummy variable equal to 1 if up for reelection, and *TEN* is years of continuous service as senator, as of 1977. All variables, however, are measured as differences between senators within each state. The numbers in parentheses are the standard errors of the estimated coefficients. The *F*-statistic implies that the null hypothesis that the entire set of included variables has no relationship with the dependent variable cannot be rejected at the .25 significance level (the critical *F*-statistic at the 0.25 level is 1.40). The sole individual regressor with a *t*-statistic greater than 1 is *CHAIR*, and the *ADA* variable has a *t*-statistic of only .852![4] The statistically significant coefficient on the political conservatism variable obtained by B-D apparently occurs because of a correlation with omitted state variables. If the B-D regression were correctly specified, our regression would also be correct and it seems unlikely that we would obtain a coefficient

[2] Of the four, only party affiliation has any simple correlation coefficient below .878 (B-D, Table 4).

[3] Interestingly enough, while differencing reduces the efficiency of the parameter estimates, the standard deviations of the key variables remain roughly equal to those in the B-D regression. For *ADA* rating, the standard deviation for individual observations is 26.14, while for within-state differences it is 30.04. For *PERCENT RETURNED*, the figures are 12.18 and 11.95, respectively.

[4] While we used the second session *ADA* ratings to be consistent with the data used in B-D, we also estimated our specification using first session ratings and obtained comparable results. Furthermore, a regression using the two-session average *ADA* ratings, which seems likely to provide a better index of political conservatism, yields an *ADA* coefficient with a *t*-statistic of .587. The equation *F*-statistic under this specification is only .929.

estimate on *ADA* ratings nearly 3 standard deviations from their estimate of +.219 (.053).[5]

To conclude, we believe that the differences obtained from the two specifications of the B-D model may be explained in any of three ways: (*i*) all variation in the transformed dependent and independent variables disappears, (*ii*) both equations are misspecified, or (*iii*) our specification is correct while the B-D specification suffers from an omitted variable problem. Furthermore, if (*iii*) is correct, then the *ADA* variable is sufficiently correlated with the state and/or allocation effects that it loses all explanatory significance when such effects are controlled for. Because (*i*) is not true, the proper explanation lies in (*ii*) or (*iii*). In either case, the estimated relationship between fiscal and political conservatism obtained by B-D is suspect.

[5] Using a Hausman (1978) specification error test which technically applies only asymptotically, we find the probability that the *ADA* parameters in the two regressions do not differ from each other, and therefore that the B-D specification is correct, to be less than .001.

REFERENCES

Bennett, James T. and DiLorenzo, Thomas J., "The Political Economy of Political Philosophy: Discretionary Spending by Senators on Staff," *American Economic Review*, December 1982, *72*, 1153–61.

Hausman, Jerry, "Specification Tests in Econometrics," *Econometrica*, November 1978, *46*, 1251–71.

The Political Economy of Political Philosophy: Reply

By JAMES T. BENNETT AND THOMAS J. DILORENZO*

In our earlier article, we developed a model which supported the hypothesis that the proportion of the staff budget which each senator from 48 states returned unspent in 1978 was significantly influenced by their political philosophy: conservatives, on average, returned a higher proportion than liberals. In their comment, Steven Cobb and Robert Hagemann emphasize that some state effects may have been omitted and, as a result, the coefficients estimated from our model may be biased. They develop an approach intended to correct this, that is, "a regression estimated on differences *between* state senators will automatically purge the regression of all state effects" (p. 524), and find that the relationship between the proportion returned and political philosophy is not statistically significant. Although concern about omitted variables is justified, there are two shortcomings with the approach that Cobb and Hagemann employ that deserve mention.

First, differencing is not an appropriate technique to use with cross-section data, as Carl Christ observed, because "the cross-section case has no analogy to the lagged value in the time series case" (1966, p. 210).[1] In a time-series, the data are uniquely ordered, but this is not true in a cross section. The data which Cobb and Hagemann difference are a cross section of two observations on senators within each of 48 states. The crux of the issue is which senator is subtracted from which within each state, because "in economics there is usually no meaningful way to order... cross-section observations" (p. 209).[2] As an illustration, consider three states (A, B, C), each with two senators (denoted by subscripts). Several differencing schemes can be proposed which generate observations for each state: (A_1-A_2),(B_1-B_2), (C_1-C_2); (A_1-A_2), (B_1-B_2), (C_2-C_1); and, (A_1-A_2), (B_2-B_1), (C_2-C_1), among others. Each scheme generates a different set of values for the dependent variable and for the independent variable measuring political philosophy. The different data sets will produce conflicting estimates of the coefficients in their model and choosing among them is necessarily arbitrary. When all permutations and combinations among 48 states are considered, Cobb and Hagemann's approach produces hundreds of different data sets that yield hundreds of conflicting estimates for the same parameters. Of course, the observations within states can be ordered by some rule (political philosophy, proportion returned unspent, tenure in office, age, etc.),[3] but the choice of a rule is itself arbitrary.

Second, for purposes of exposition, consider the case where senators are either extreme liberals or conservatives; liberals spend all of their staff budgets, whereas conservatives spend none of their funds; also, every state has two senators with the same political philosophy. In this ideal case, our model will verify that political philosophy determines spending behavior, but the model presented by Cobb and Hagemann will not, for when differences are taken, the value of the dependent variable (and the independent variable for political philosophy) for every observation is identically zero and no empirical estimates can be obtained whatsoever. One

*George Mason University, Fairfax, VA 22030. Research support provided by the Sarah Scaife Foundation and the Earhart Foundation is gratefully acknowledged.

[1] In a strict sense, Cobb and Hagemann are differencing, not employing "lagged values." However, Christ explicitly points out that "...using first differences as variables is equivalent to using lagged variables" (p. 177).

[2] Christ adds that "In a pure cross-section model... the observations typically have no natural order, though in certain cases we can *imagine* putting them in the order of size, social status, or distance from some focal point, or what not" (p. 209, emphasis added).

[3] If political philosophy is used, the question then becomes which measure of political philosophy. We used three: the American Conservative Union, the AFL-CIO, and the Americans for Democratic Action rankings of senators.

must doubt the usefulness of any approach that does not empirically confirm a known fact.

In sum, we question any approach that 1) produces numerous conflicting estimates of the model's parameters with only arbitrary rules for choosing among them, and 2) fails to confirm empirically a relationship known to be valid a priori.

REFERENCES

Christ, Carl F., *Econometric Methods and Models*, New York: Wiley & Sons, 1966.

Cobb, Steven and Hagemann, Robert P., "The Political Economy of Political Philosophy: Comment," *American Economic Review*, June 1984, *74*, 523–25.

The Separability of Production and Location Decisions: Comment

By HARRY R. CLARKE*

In a recent note in this *Review*, Arthur Hurter, Joseph Martinich, and Enrique Venta (hereafter Hurter et al.) consider the question of the conditions under which the location decision for a cost-minimizing firm can be treated separately from the decision governing the firm's desired input mix. The authors argue that if the firm's production function is homothetic, total costs can be minimized by first determining an optimal location for the facility, and then determining the optimal input mix at this location. The implication of this viewpoint is that facility location and facility design problems can be analyzed independently. Thus the traditional theory of the firms which abstracts from spatial elements, and the traditional Weberian theory of industrial location which does not consider the choice of input mix, remain independently valid without involving the possibility of suboptimal decision making on the part of the firm. My purpose here is to clarify the conditions under which this type of separability does obtain: in particular, I argue that the facility location problem can be solved independently of the input mix selection problem if and only if a firm's production function is of the fixed proportions type. This implies that homotheticity is neither necessary nor sufficient for separability of these decision problems.

I. The Model

To facilitate comparison with the earlier analysis, I follow the assumptions and notation of Hurter et al. The problem is that of choosing a facility location $x=(x_1,x_2)$ and an input mix vector $z=(z_1,z_2,\dots z_n)$ so as to minimize production and transportation costs while producing output at a fixed rate $\bar{z}_0$. This optimization task takes the form:

$$\text{(P1)} \qquad \min_{x,z} \sum_{j=1}^{n} \left[r_j d(x,a_j) + p_j \right] z_j,$$

$$\text{subject to} \quad \bar{z}_0 = f(z_1, z_2, \dots z_n),$$

where r_j = transportation cost of with quantity of factor j per unit distance, $d(x,a_j)$ = an l_p measure of distance from location x to the source of input $j, a_j=(a_{1j},a_{2j})$, p_j = price per unit of input j at its source, and $f(z_1,\dots z_n)$ = firm's production function.

Hurter et al. assume f to be differentiable so for an arbitrary location $x=\bar{x}$, first-order conditions for minimization of cost over z require

$$\text{(1)} \qquad r_j d(\bar{x},a_j) + p_j - \mu(\bar{x}) f_j(z^*) = 0, \qquad j=1,2,\dots n;$$

$$\text{(2)} \qquad \bar{z}_0 = f(z_1^*, z_2^*, \dots z_n^*) = f(z^*),$$

where $\mu(\bar{x})$ is the Lagrange multiplier associated with location $\bar{x}$ and where the starred superscript denotes the optimizing value of a variable.

Multiplying the jth first-order condition in (1) by z_j^* and summing over j, we have

$$\text{(3)} \qquad \sum_{j=1}^{n} \left[r_j d(\bar{x},a_j) + p_j \right] z_j^* = \mu(\bar{x}) \sum_{j=1}^{n} z_j^* f_j(z^*).$$

Hurter et al. then note that when output is homothetic and output is held fixed:

$$\text{(4)} \qquad \sum_{j=1}^{n} z_j^* f_j(z^*) = K > 0,$$

a constant.

[*] Industrial Engineering and Management Division, Asian Institute of Technology, G.P.O. Box 2754, Bangkok, Thailand 10501. I thank O. Fujiwara, A. P. Hurter, Jr., and R. M. Shrestha for helpful remarks on earlier versions. I alone am responsible for views expressed.

Substituting this result in (3) suggests replacing the constrained optimization problem (P1) with the unconstrained task:

$$(5) \qquad \min_{\bar{x}} K\mu(\bar{x}) \Leftrightarrow \min_{x} \mu(x).$$

Hurter et al. deduce from this that if the production function f is homothetic, (P1) can be solved by *first* determining an optimal location x^* and, given this location, by *then* determining the optimal input mix z^* at this location (compare their Theorem 1).

This result does not, however, follow from the preceding argument. In particular, I now show homotheticity is not sufficient to guarantee separability of the type indicated. First note that apart from (1), an additional first-order condition for optimality is the constraint (2). There are thus in general $(n+1)$ independent first-order conditions for optimality. Assuming these conditions do in fact describe a unique global minimum (a sufficient condition here is that f be strictly regular quasi concave), these equations determine the $(n+1)$ variables z_1^*, $z_2^*, \dots z_n^*$, $\mu(\bar{x})$ given the site $\bar{x}$ and the parameters r_j, p_j and $a_j (j=1,2,\dots n)$. Thus $\mu(\bar{x})$, which Hurter et al. correctly identify as the marginal cost of production at location $\bar{x}$ is determined *simultaneously* with the desired input levels $z_1^*, z_2^*, \dots z_n^*$. Formally by rewriting (3) as

$$(6) \qquad \mu(\bar{x}) = \sum_{j=1}^{n} \left[r_j d(\bar{x}, a_j) + p_j\right] z_j^* \Big/ \sum_{j=1}^{n} z_j^* f_j(z^*),$$

we see that even if the production function is homothetic, so that the denominator of the right-hand side of (6) is constant, that $\mu(\bar{x})$ is still determined *jointly* with $z_1^*, z_2^* \dots z_n^*$ for a given location $\bar{x}$. The only condition which eliminates this type of simultaneity is prior knowledge that factor input levels are predetermined by a fixed proportions production technology. In this event there is no input mix or "design" problem since factor proportions are predetermined at all locations. Thus the only problem facing planners is the selection of a location which minimizes the transportation costs of inputs in accord with the traditional Weberian analysis. In particular, since homotheticity does not guarantee fixed proportions in production, it is not sufficient to ensure the type of separability envisaged by Hurter et al.

This argument does not imply there is anything incorrect about writing (P1) in form (5). It simply demonstrates the intuitively reasonable conclusion that, if factor substitution is possible, input selection will depend on the transport costs of inputs. In the same way the choice of location depends on the desired input mix: *ceteris paribus*, a facility will be located relatively close to "cheap" inputs. Only the assumption of fixed proportions in production eliminates this type of simultaneity.

On the other hand, it is by no means necessary for f to be homothetic for the criterion function in (P1) to be representable in a form that depends only on the location variable $\bar{x}$ and the parameters of the production function. With appropriate restrictions on the Hessian of f, equations (1) and (2) can be solved locally for the $(n+1)$ variables $z_1^*, \dots z_n^*$, and $\mu(\bar{x})$ for a given location $\bar{x}$ and for given price and production function parameters. This again does not imply any recursive separability in the location and design problems, since in transforming the problem to a form that depends only on the location variable, one must necessarily use information on the relation between optimal input levels and location. For this reason it does not seem worthwhile to carry this local analysis further and to consider, for example, the conditions under which this type of local analysis could be globalized.

II. Conclusions and Final Remarks

In general, input mix and facility location problems are not separable for a cost-minimizing firm even if the firm's production function is homothetic. In fact, necessary and sufficient conditions for separability are that production technologies are of the fixed-proportions type when the question of input mix selection is predetermined in any event. If the firm's technology is homothetic, then its criterion function can certainly be expressed in terms of locational variables alone. The practical problem with utilizing such a procedure is that the evaluation of such a function involves the simultaneous computation of desired input levels at each location. Locally such a procedure can be carried out with much weaker assumptions than homotheticity, although this procedure is again not helpful, since it does not imply any recursive or sequential separability in the structure of the decision problem.

There are further aspects of the Hurter et al. analysis that deserve at least brief comment. David Emerson (1973) has shown that "separability" of the type discussed by Hurter et al. does not arise when there are quantity discounts in the transportation of final output to demand centers. R. M. Shrestha and I (1982) have shown, in the context of an energy facility planning problem, that output wastage ("line losses") in distribution again destroys separability *even if* production technologies are of the fixed proportions type.

The real strength of the homotheticity assumption in production-location theory rests on the implied independence of desired location from planned output levels. If homotheticity is strengthened to constant returns to scale, then this invariance persists even in the presence of proportional product distribution costs. An interesting application of these ideas, in a Noboru Sakashita (1980) type linear space model, is the analysis of optimal siting under demand uncertainty by Chao-cheng Mai (1981). Mai shows that, with constant returns to scale, optimal location is independent of the extent of demand uncertainty.

REFERENCES

Clarke, H. R. and Shrestha, R. M., "Location and Input Mix Decisions for Energy Facilities," paper prepared for the Asia Pacific Conference on Operational Research, National University of Singapore, November, 1982.

Emerson, D. L., "Optimal Firm Location and the Theory of Production," *Journal of Regional Science*, December 1973, *13*, 335–45.

Hurter, A. P., Jr., Martinich, J. S., and Venta, E. R., "A Note on the Separability of Production and Location," *American Economic Review*, December 1980, *70*, 1042–45.

Mai, C., "Optimum Location and the Theory of the Firm Under Demand Uncertainty," *Regional Science and Urban Economics*, December 1981, *11*, 549–57.

Sakashita, N., "The Location Theory of the Firm Revisited: Impacts of Rising Energy Prices," *Regional Science and Urban Economics*, December 1980, *10*, 423–28.

The Separability of Production and Location Decisions: Reply

By Arthur P. Hurter, Jr., Joseph S. Martinich and Enrique R. Venta*

We welcome this opportunity to clarify some aspects of our paper and to respond to Harry Clarke's comment. Clarke appears to make three major points: 1) the production (input mix) and location subproblems of our equation (1) (his (P1)) are *independent* only when the production function has the property of fixed proportions; 2) a *separation* of the production and location variables can be obtained with weaker assumptions about the production function than homotheticity; and 3) while recognizing, as have we and others, that homotheticity of the production function means that the same location will be optimal for all levels of output, homotheticity provides no additional benefit or simplification.

That we generally agree with the first of Clarke's points is evident from some of our other publications (see Hurter and Venta, 1982; Hurter and R. E. Wendell, 1972; Martinich and Hurter, 1982). In general, we also concur with Clarke's second point. However, we feel that he has missed some aspects of these first two points which led him to his third point with which we do not agree.

Certainly, the determination of the optimal input mix and optimal location are not independent except in cases where the production function exhibits fixed proportions. Unfortunately, on page 1045 of our paper, we stated that with homotheticity there was *independence* between the location and input variables when we meant that a kind of *separation* was possible as illustrated by Hurter and Wendell. We intended to imply that a homothetic production function facilitates writing the constrained optimization problem (P1) as a problem explicitly in the location variable *only*, which contains all the relevant information on input substitution. Solving this problem *does* yield the optimal location, and *subsequent* substitution into the first-order conditions does yield the optimal input mix.

While the location and input mix problems are not independent, with the homotheticity assumption the two problems can easily be separated and the optimal location found without *explicitly* solving for the production variables. With the exception of the reference to independence noted earlier, throughout our paper we referred to this "separation" of the location and input variables intending the same meaning as that used in earlier papers by Hurter and Wendell and by A. J. Goldman (1974).

With respect to Clarke's second point, assuming the usual second-order conditions, the first-order conditions can be solved locally to provide a form of cost function in terms of the parameters of the problem and the location variables. Thus, a separability exists, locally at least, between the location variables and the input variables. Although in theory this can be done for general production functions, in practice this requires solving $n+1$ possibly nonlinear equations in $n+1$ unknowns, which is a difficult and sometimes intractable task. Homotheticity of the production function allows one to obtain a *global closed-form* representation of marginal cost as a function of location without solving a complex system of equations. This reduces the original $n+2$ variable problem to a simpler two-variable problem in only the location variables. The explicit separation, with homothetic production functions, has been performed for Cobb-Douglas production functions by Hurter and Wendell, and for the *CES* production function and various two-stage production processes by Goldman.

*Hurter: Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60201-9990; Martinich: School of Business Administration, University of Missouri, St. Louis, MO 63121; Venta: Department of Management Science, Loyola University, Chicago, IL 60611.

Furthermore, Goldman has shown that the location problem resulting from this separation can be solved using a slightly modified version of the common Weber problem algorithms.

Consequently, we cannot agree that the assumption of a homothetic production function serves only to provide independence of the optimal location from the planned output level. We agree that homotheticity does lead to the same optimal location for all output levels and stated this in Theorem 2 of our earlier paper. However, we feel that the kind of separation provided by the homotheticity, as described above and included in our Theorem 1, not only can simplify solution of the production-location problem, but that furthermore only slight modification of standard algorithms is required for solution.

As a final point, the role of homotheticity in problems wherein the cost of transporting output and wherein the transport costs and input prices depend on the quantities shipped and purchased is addressed by Hurter and Venta. We believe that paper adequately covers the points made in the penultimate paragraph of Clarke's comment.

REFERENCES

Clarke, H. R., "The Separability of Production and Location Decisions: Comment," *American Economic Review*, June 1984, *74*, 528–30.

Goldman, A. J., "Fixed-Point Solution of Plant Input/Location Problems," *National Bureau of Standards Journal of Research — B Mathematical Sciences*, April/June 1974, *78B*, 79–94.

Hurter, A. P., Jr., Martinich, J. S. and Venta, E. R., "A Note on the Separability of Production and Location," *American Economic Review*, December 1980, *70*, 1042–45.

______**and Venta, E. R.**, "Production-Location Problems," *Naval Research Logistics Quarterly*, June 1982, *29*, 279–90.

______**and Wendell, R. E.**, "Location and Production—A Special Case," *Journal of Regional Science*, August 1972, *12*, 243–47.

Martinich, J. S. and Hurter, A. P., Jr., "Price Uncertainty and the Optimal Production-Location Decision," *Regional Science and Urban Economics*, November 1982, *12*, 509–28.

Migration and Asymmetric Information: Comment

By ELIAKIM KATZ AND ODED STARK*

In a recent issue of this *Review*, Viem Kwok and Hayne Leland (hereafter K-L) offer a novel explanation of the loss to *LDC*s of their foreign-trained, highly skilled labor—the so-called "brain drain": asymmetric information in the labor market. Employers abroad—where the skills are acquired—have better information than home-country employers with respect to worker productivity. Hence, foreign employers, being able to (more) accurately observe the productivity of a potential employee, match that productivity with an appropriate wage, whereas domestic employers, who cannot, offer returning workers a wage equal to the *average productivity* of the whole group of returning workers. Hence, the better employees will prefer not to return, thus generating the brain drain phenomenon. This holds true even when a given worker's productivity is the same in both countries, and when, for equal wages, workers prefer home to abroad.

It is our purpose in this comment to show that K-L's treatment of migration with asymmetric information is only one of a number of possible scenarios. We demonstrate that under alternative (and in our opinion more likely) scenarios of asymmetric information, less-skilled workers who would have stayed in their home country had information been perfect, migrate to the rich country. Thus, asymmetric information may bring about the migration from *LDC*s of the less-skilled workers rather than of the most skilled.

Let $W_{US}(\theta)$ and $W_T(\theta)$ be the wage paid to workers whose skill level is θ, in the (rich) United States (US) and (poorer) Taiwan (T), respectively. Assume that Taiwanese workers regularly apply a discount factor to U.S. wages when comparing them to Taiwanese wages because of preference for Taiwanese lifestyle, culture, etc. When making the migration decision, they compare $kW_{US}(\theta)$ to $W_T(\theta)$ where $k<1$. Thus, a worker will wish to migrate from Taiwan to the United States if $W_T(\theta) < kW_{US}(\theta)$.

Let us begin the analysis by assuming that in a given occupation the (positive) wage differential between the United States and Taiwan does not vary with θ. Furthermore, let us assume, without loss of generality, that θ is distributed on $[0,1]$ with distribution functions $F_{US}(\theta)$ and $F_T(\theta)$, respectively.

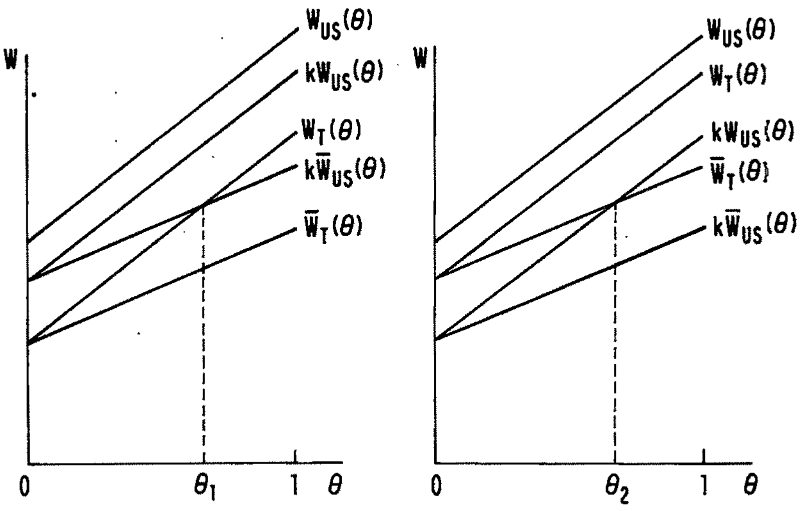
Two cases may now be distinguished. First, the difference between the U.S. and Taiwanese wage may be sufficiently large so that the $kW_{US}(\theta)$ line lies above the $W_T(\theta)$ line. This situation is depicted in Figure 1, panel (a). In this case, all members of the occupation will wish to migrate from Taiwan to the United States in the presence of full information.

Alternatively, if the difference between the U.S. and Taiwanese wage is small, the $kW_{US}(\theta)$ line will lie below the $W_T(\theta)$ line. In this situation, as depicted in Figure 1, panel (b), under full information no members of the occupation will wish to migrate to the United States.

Let us now consider the effect of information asymmetry on the migration decision in each of the above cases. The effect may be seen to crucially depend on the location of the ignorant employers. First consider what happens if, as seems very likely for most occupations, U.S. employers are unable to observe the individual θ of Taiwanese immigrants. In this case, U.S. employers will pay each Taiwanese worker the average product of the Taiwanese immigrants. Calling this wage $\overline{W}_{US}(\theta)$, it is clear that $\overline{W}_{US}(0) = W_{US}(0)$ whereas when $\theta > 0$ $\overline{W}_{US}(\theta) < W_{US}(\theta)$. Hence, the $k\overline{W}_{US}(\theta)$ line is as depicted in Figure 1(a).

It is immediately clear from Figure 1(a) that, whereas in the presence of full information, all members of the occupation will wish to migrate to the United States, once asym-

*Bar-Ilan University, Ramat-Gan, Israel, and Bar-Ilan University and Harvard University, Cambridge, MA 02138, respectively.



(a) (b)

FIGURE 1

metric information is incorporated into the model, *only the less-skilled workers* in the $[0, \theta_1]$ interval will wish to migrate.

The occupation considered may, of course, be one of the very few where foreign (U.S.) employers have the better information because the skills required may only be acquired in the United States. In such a case, Taiwanese employers will pay each worker the average wage of his group. By reasoning similar to the above, this mean wage line, $\overline{W}_T(\theta)$ is as drawn in Figure 1(a) and hence, in this scenario, *all* members of the occupation (and not only the most skilled) will wish to migrate (fail to return).

Turning to the second case as depicted in Figure 1(b), we carry out the same analysis. This produces, in the presence of U.S. employers being ignorant, the mean wage to Taiwanese immigrants of $k\overline{W}_{US}(\theta)$, and in the presence of Taiwanese employers being ignorant the mean wage to Taiwanese returnees of $\overline{W}_T(\theta)$. It is clear from the diagram, therefore, that in this case, the migration pattern under asymmetric information will be such that if the lack of information is in the United States, no migration will take place, and if it is in Taiwan, migration of the top skilled workers $[\theta_2, 1]$ will occur. This last case corresponds to K-L's analysis.

If we now remove the assumption of the constancy of the wage differential in a given occupation, our results will be amplified or modified depending on whether the differential increases or decreases with θ. The essence of our results, however, is in the asymmetry of information and its location. If the information is with Taiwanese employers (and this seems the more common situation), asymmetric information will, if it has any effect, tend to encourage migration of low-skilled members of the occupation. Only if the information about Taiwanese workers is with U.S. employers (but not with Taiwanese employers) and then only under restrictive conditions, may asymmetric information cause a migration pattern skewed towards the top-skilled workers.[1]

[1] For example, it is easily demonstrated using our diagrammatic exposition (and this is left to the reader), that even with asymmetric information at T and further, even with an increasing intercountry wage differential, the migration (from T to US) is characterized by a lower *average*-skill level as some *less*-skilled workers, who would have had no wish to migrate had information at T been perfect, find migration attractive when information at T is imperfect.

REFERENCE

Kwok, Viem and Leland, Hayne, "An Economic Model of the Brain Drain," *American Economic Review*, March 1982, *72*, 91–100.

Migration and Asymmetric Information: Reply

By PETER KWOK AND HAYNE LELAND*

In our 1982 article, we showed that more talented workers will migrate to the country where their talents are recognized, even when wages are "fair" in both countries. We focused on the most skilled set of workers—those associated with the "brain drain" problem. If the country of education of these persons (assumed to be the United States in our example) can better screen their talents than the country of origin (assumed to be Taiwan), then a brain drain problem may well exist.

In their comment, Eliakim Katz and Oded Stark provide another example of our result. If the information asymmetries are reversed (i.e., Taiwan can better screen Taiwanese workers than the United States), then it follows immediately from our analysis (reverse the subscripts!) that the opposite result holds: talented workers will remain at home and less talented workers may well migrate. It should be noted that the coefficient k must now be interpreted as the ratio of Taiwanese wages to U.S. wages, divided by the original k which reflected the preference of Taiwanese for remaining in Taiwan. Since this original k was less than one, the modified k can be less than one (consistent with migration of the less talented—see our Proposition 2) only if U.S. wages exceed Taiwanese wages by a sufficient amount. Thus, in contrast with our analysis which admits emigration even when wages are equal for workers of equal productivity, emigration of the least skilled will occur only when there is a sufficient wage differential—and even then, emigration will often be partial (see our Proposition 6).

We certainly don't argue with the Katz-Stark result, since it follows directly from our analysis. But it is most likely true that the pool of workers for which their conclusion follows is not the highly educated (and often foreign-trained) group associated with the original brain drain problem. On the other hand, the direction of asymmetry assumed by Katz and Stark may characterize a larger set of workers, and may help explain the relative immobility of labor even in the presence of substantial wage differentials. For as suggested above, emigration of the least skilled will occur only when the wage differential is sufficiently large to overcome the "stickiness" created by informational asymmetry, and even then only a fraction of workers will move.

It is clear that informational asymmetries play an important role in explaining the mobility of labor. The analysis by Katz and Stark confirms our conclusion that informational aspects as well as relative wages must be included in any discussion of labor mobility and in any assessment of government policies affecting mobility.

REFERENCES

Katz, Eliakim and Stark, Oded, "Migration and Asymmetric Information: Comment," *American Economic Review*, June 1984, *74*, 533–34.

Kwok, Viem (Peter) and Leland, Hayne, "An Economic Model of the Brain Drain," *American Economic Review*, March 1982, *72*, 91–100.

*Crocker National Bank, Hong Kong, and University of California, Schools of Business Administration, 350 Barrows Hall, Berkeley, CA 94720, respectively.

Auditors' Report

February 27, 1984

Executive Committee
The American Economic Association

We have examined the balance sheets of the American Economic Association as of December 31, 1983 and 1982 and the related statements of revenues and expenses, changes in general fund and restricted fund balances and changes in financial position for the years then ended. Our examinations were made in accordance with generally accepted auditing standards and, accordingly, included such tests of the accounting records and such other auditing procedures as we considered necessary in the circumstances.

In our opinion, the financial statements referred to above present fairly the financial position of the American Economic Association as of December 31, 1983 and 1982, its revenues and expenses and the changes in its financial position for the years then ended, in conformity with generally accepted accounting principles applied on a consistent basis.

Touche Ross and Co.
Certified Public Accountants
Nashville, Tennessee

The American Economic Association Balance Sheets, December 31, 1983 and 1982

	1983	1982
Assets		
Cash	$ 705,732	$ 640,267
Investments, at market (Notes A and B)	3,227,487	2,808,085
Accounts Receivable, less allowance for doubtful accounts of $1,362 (1983) and $337 (1982)	174,787	76,375
Inventory of *Index of Economic Articles*, at cost	47,992	37,511
Prepaid Expenses	21,287	15,022
Office Furniture and Equipment, at cost, less accumulated depreciation of $21,010 (1983) and $17,622 (1982)	43,634	27,276
	$4,220,919	$3,604,536
Liabilities and Fund Balances		
Accounts Payable and Accrued Liabilities	$ 342,639	$ 361,952
Deferred Revenue (Note A):		
Life membership dues	49,680	52,302
Other membership dues	476,663	479,474
Subscriptions	401,254	415,751
Job Openings for Economists	17,004	16,986
	944,601	964,513
Accrual for Directory (Note A)	147,141	87,141
Fund Balances:		
Restricted	88,304	42,636
General	2,468,490	2,110,632
Unrecognized change in market value of investments (Notes A and C)	229,744	37,662
Net Worth	2,698,234	2,148,294
Total Fund Balances	2,786,538	2,190,930
	$4,220,919	$3,604,536

See notes to financial statements.

The American Economic Association Statements of Revenues and Expenses for the Years Ended December 31, 1983 and 1982

	1983	1982
Revenues from Dues and Activities:		
Membership dues and subscriptions	$757,999	$729,492
Nonmember subscriptions	630,951	645,362
Job Openings for Economists subscriptions	27,247	24,980
Advertising	98,454	87,350
Sale of *Index of Economic Articles*	61,160	64,171
Sale of copies, republications, and handbooks	27,134	31,813
Sale of mailing list	34,517	38,215
Annual meeting	34,033	36,854
Sundry	51,365	41,428
	1,722,860	1,699,665
Investment Gains (Note B)	165,247	329,584
Net Revenues	**1,888,107**	**2,029,249**
Publication Expenses:		
American Economic Review	480,228	451,591
Journal of Economic Literature	637,573	626,065
Directory publication (Note A)	60,000	55,000
Job Openings for Economists	49,754	46,029
Index of Economic Articles	32,958	28,712
	1,260,513	1,207,397
Operating and Administrative Expenses:		
General and administrative:		
Salaries	161,208	141,980
Rent	13,282	12,016
Other (Exhibit I)	156,401	156,641
Committee	44,585	49,304
Annual meeting	4,760	5,400
Provision for federal income taxes (Note A)	2,000	2,500
	382,236	367,841
Total Expenses	**1,642,749**	**1,575,238**
Revenues in Excess of Expenses	$ **245,358**	$ **454,011**

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN GENERAL FUND BALANCE

	Total	Operations	Market Value Adjustments
Balance at January 1, 1982	**$1,569,287**	**$921,616**	**$647,671**
Add market value adjustments resulting from inflation (Note A)	87,334	–	87,334
Add revenues in excess of expenses	454,011	454,011	–
Balance at December 31, 1982	**2,110,632**	**1,375,627**	**735,005**
Add market value adjustments resulting from inflation (Note A)	112,500	–	112,500
Add revenues in excess of expenses	245,358	245,358	–
Balance at December 31, 1983	**$2,468,490**	**$1,620,985**	**$847,505**

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN RESTRICTED FUND BALANCE

	Balance at January 1	Receipts	Disbursements	Balance at December 31
YEAR ENDED DECEMBER 31, 1982:				
The Alfred P. Sloan Foundation and Federal Reserve System grants for increase of educational opportunities for minority students in economics	$10,500	$114,500	$123,090	$ 1,910
The Minority Scholarship fund for minority students applying for graduate work in economics	5,000	–	–	5,000
The Rockefeller Foundation grant for minority students applying for graduate work in economics	–	58,500	25,105	33,395
Sundry	2,231	100	–	2,331
	$17,731	$173,100	$148,195	$42,636
YEAR ENDED DECEMBER 31, 1983:				
The Alfred P. Sloan Foundation and Federal Reserve System grants for increase of educational opportunities for minority students in economics	$ 1,910	$120,500	$ 95,650	$ 26,760
The Minority Scholarship fund for minority students applying for graduate work in economics	5,000	–	–	5,000
The Rockefeller Foundation grant for minority students applying for graduate work in economics	33,395	58,500	40,115	51,780
Sundry	2,331	5,300	2,867	4,764
	$42,636	$184,300	$138,632	$88,304

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN FINANCIAL POSITION FOR THE YEARS ENDED DECEMBER 31, 1983 AND 1982

	1983	1982
Cash, beginning of year	$640,267	$557,710
SOURCES OF CASH:		
Revenues in excess of expenses	245,358	454,011
Noncash charges:		
Depreciation	3,677	2,609
Directory publication (Note A)	60,000	55,000
Market value adjustments (Note A)	57,572	(120,654)
Cash provided by operations	366,607	390,966
INCREASE (DECREASE) IN CASH DUE TO CHANGES IN:		
Investments	(419,402)	(896,738)
Accounts receivable	(98,412)	5,220
Inventory of *Index of Economic Articles*	(10,481)	28,120
Prepaid expenses	(6,265)	43
Office furniture and equipment	(20,035)	(6,601)
Accounts payable and accrued liabilities	(19,313)	76,069
Deferred revenue	(19,912)	37,135
Accrual for directory	–	(1,177)
Restricted funds	45,668	24,905
General fund, market value adjustment	112,500	87,334
Unrecognized change in market value of investments	134,510	337,281
Cash, end of year	$705,732	$640,267

See notes to financial statements.

Notes to Financial Statements

A. Summary of Significant Accounting Policies

Investments are accounted for on a market value basis. According to the method the Association uses to value investments, the change in market value of corporate stocks, government obligations, bonds and commercial paper during the year, after adjusting for an inflation factor (4.2% in 1983 and 4.6% in 1982), is recognized in income over a three-year period for corporate stocks and reflected in current income for government obligations, bonds and commercial paper.

The accrual for directory results because every three to five years the Association publishes a directory which lists, among other things, the names and addresses of its membership. This directory was most recently published in 1981 and distributed at no cost to the membership. In order to properly match the publishing cost of this directory with revenue from membership dues, the Association provided $60,000 in 1983 and $55,000 in 1982 for estimated publishing costs which will reduce actual directory expenses in the year of publication.

Deferred revenue represents income from membership dues and subscriptions to the various periodicals of the Association which are deferred when received. These amounts are then recognized as income following the distribution of the specified publications to the members and subscribers of the Association. Income from life membership dues is recognized over the estimated average life of these members.

The American Economic Association files its federal income tax return as an educational organization, substantially exempt from income tax under Section 501(c) (3) of the Internal Revenue Code. As required by Section 511(a) of this Code, the Association provides for federal income taxes on certain revenues which are not substantially related to its tax exempt purpose. This "unrelated business income" includes income from advertising and the sale of mailing lists. The Association has been determined to be an organization which is not a private foundation.

B. Investments and Investment Income

Investments consist of:

	December 31, 1983		December 31, 1982	
	Cost	Market	Cost	Market
Government obligations, bonds and commercial paper	$ 847,212	$ 908,212	$ 829,121	$ 927,110
Corporate stocks and mutual funds	1,649,062	2,319,275	1,237,507	1,880,975
	$2,496,274	$3,227,487	$2,066,628	$2,808,085

Investment gains (losses) recognized consist of:

	Year Ended December 31	
	1983	1982
Government obligations, bonds, and commercial paper:		
Interest	$151,144	$146,549
Increase (decrease) in market value recognized	(76,584)	72,091
	74,560	218,640
Corporate stocks and mutual funds:		
Cash dividends	71,675	62,381
Increase in market value recognized (Note C)	19,012	48,563
	90,687	110,944
Investment gains, net	$165,247	$329,584

C. Unrecognized Change in Market Value of Investments

As described more fully in Note A, the Association recognizes in income over a three-year period changes in the market value of its corporate stocks. The following summarizes the years in which market value changes in stocks occurred that affected 1983 and 1982 revenues, and the amount of these market value increases (decreases) that will be recognized in income in future periods.

Year of Market Value Change	Recognized in Income in		To be Recognized in		Unrecognized Change	
	1983	1982	1984	1985	1983	1982
1980	$ –	$ 99,298	$ –	$ –	$ –	$ –
1981	(139,131)	(139,131)	–	–	–	(139,131)
1982	88,396	88,396	88,397	–	88,397	176,793
1983	69,747	–	69,747	71,600	141,347	–
	$ 19,012	$ 48,563	$158,144	$71,600	$229,744	$ 37,662

The Association's revenues in excess of expenses would have been $437,440 in 1983 and $670,639 in 1982 if changes in the market value of corporate stocks, after adjustment for inflation, had been recognized only, but entirely, in the year in which they occurred.

D. Retirement Annuity Plan

Employees of the Association are eligible for participation in a contributory retirement annuity plan. Payments by the Association and participating employees are based on the employee's compensation. Benefit payments are based on the amounts accumulated from such contributions. The total pension expense was approximately $23,000 and $21,000 for 1983 and 1982, respectively.

E. Ratio of Net Worth to Expenses

The ratio of net worth at December 31, 1983 to 1984 budgeted expenses is 1.48 and the ratio of net worth at December 31, 1982 to actual 1983 expenses is 1.31.

EXHIBIT 1—THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF OTHER GENERAL AND ADMINISTRATIVE EXPENSES FOR THE YEARS ENDED DECEMBER 31, 1983 AND 1982

	1983	1982
Dues and subscriptions	$ 39,925	$ 42,405
Mailing list file maintenance	24,385	20,535
Postage	15,788	19,751
Periodic mailing expenses	14,118	14,019
Accounting and legal	12,750	12,200
Investment counsel and custodian fees	11,267	8,474
Office supplies	10,582	11,271
Insurance and miscellaneous	9,006	4,062
President and president-elect expenses	5,315	4,368
Telephone	5,033	4,010
Depreciation (straight-line method)	3,677	2,609
Currency exchange charges	2,160	1,614
Uncollectible receivables	1,658	540
Travel and entertainment	737	783
Economic Institute Fellowships	–	10,000
	$156,401	$156,641

NOTES

1985 Nominating Committee of AEA

In accordance with Section IV, paragraph 2, of the bylaws of the American Economic Association as amended in 1972, President-Elect Charles P. Kindleberger has appointed a Nominating Committee for 1985 consisting of Gardner Ackley, Chair; Padma Desai, James A. Heffner, Ronald W. Jones, Stanley Lebergott, Martin F. Prachowny, and Thomas E. Weiskopf.

Attention of members is called to the part of the bylaw reading, "In addition to appointees chosen by the President-Elect, the Committee shall include any other member of the Association nominated by petition including signatures and addresses of not less than 2 percent of the members of the Association delivered to the Secretary before December 1. No member of the Association may validly petition for more than one nominee for the Committee. The names of the Committee shall be announced to the membership immediately following its appointment and the membership invited to suggest nominees for the various officers to the Committee."

Nominations for AEA Officers: 1985

The Electoral College on March 23 chose Alice M. Rivlin as nominee for President-Elect of the American Economic Association in the balloting to be held in the autumn of 1984. Other nominees (chosen by the 1984 Nominating Committee) are: Vice President (two to be elected), Elizabeth E. Bailey, Gerard Debreu, Thomas C. Schelling, and Joseph E. Stiglitz, for members of the Executive Committee (two to be elected), Marcus Alexis, Alan S. Blinder, Daniel McFadden, and Sherwin Rosen.

Under a change in the bylaws as described in the *American Economic Review Proceedings*, May 1971, page 472, additional candidates may be nominated by petition, delivered to the Secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Association for the office of President-Elect, and not less than 4 percent for each of the other offices. For the purpose of circulating petitions, address labels will be made available by the Secretary at cost.

The Executive Committee of the Association, meeting on December 27, 1983 in San Francisco, voted to establish a Search and Structure Committee for the *Journal of Economic Literature*. It was asked to recommend a replacement for the current Managing Editor, Moses Abramovitz, who has indicated that he does not wish to serve beyond his current term, which expires at the end of 1985. The Committee was also asked to recommend with respect to the possible reintegration in a single place of editorial functions that, since 1981, have been divided between Stanford and Pittsburgh.

President Charles L. Schultze has now appointed such a Committee, with the following members: Gardner Ackley (Chair), James Buchanan, Alan Blinder, Albert Rees, Kerry Smith, Stanley Black, and Allen Kelley.

The Committee urgently invites formal or informal communications from members, proposing themselves, or recommending others, to serve for a three-year renewable term as Managing Editor of the *JEL*, beginning January 1, 1986. The date of transfer of responsibilities to the new Editor can be negotiated. Communications may be sent to the Secretary of the Association; to Gardner Ackley, Department of Economics, University of Michigan, Ann Arbor, MI 48109; or to another member of the Committee. Such communications will be treated as confidential, if so requested. It is hoped that appointment of the new Editor can be officially accomplished at the Executive Committee Meeting of December 1984, or (at the latest) of March 1985.

The Midwest Economic Association is seeking applicants for the office of Secretary-Treasurer. Assurance of support by the candidate's institution is desirable. Those interested should contact Rendigs Fels, MEA President, Box 1664, Station B, Vanderbilt University, Nashville, TN 37235.

The Department of Health and Human Services announces the publication of the *Final Report of the Seattle-Denver Income Maintenance Experiment* (SIME/DIME). The three volumes are the last in a series of four large-scale income maintenance experiments of the late 1960's and 1970's. For a free copy of any or all, write to SIME/DIME Distribution Center, DHHS/ASPE/Office of Income Security Policy, Rm 410E Humphrey Bldg, 200 Independence Avenue, SW, Washington, D.C. 20201.

The National Council for Soviet and East European Research invites proposals for research contracts in its annual competitions with deadlines of November 1 each year. The Council is a nonprofit educational corporation which conducts national programs of research dealing with major policy issues and questions of Soviet and East European social, political, and historical development. Eligibility is limited to scholars at the postdoctoral level or with an equivalent degree of professional maturity. Council contracts are awarded through U.S. universities for collaborative or individual projects no longer than two years in duration. For further information, write The National Council for Soviet and East European Research, 1755 Massachusetts Avenue, NW, Suite 304, Washington, D.C. 20036. (Telephone 202+387-0168)

The National Science Foundation's Division of Policy Research and Analysis (PRA) announces a program to support longer-term projects to address continuing federal policy issues or to develop improved methods for policy analysis, and concept papers to develop new ideas for improving policy analysis. The PRA program is divided into three subject areas: policy science; technology and resource policy; science and innovation policy. For additional information, contact Dr. Peter W. House, Director, Division of Policy Research and Analysis, National Science Foundation, Washington, D.C. 20550.

The National Institute of Aging invites grant applications for research projects designed to extend scientific understanding of how and why particular variations in the social context may have positive or negative effects, and to identify influences that may improve the health or effective functioning of middle-aged and older people. The deadlines are March 1, July 1, and November 1. For program announcements, see *NIH Guide on Grants and Contracts*, or write National Institute on Aging (Social Environments), Bldg. .31C, Rm 4C32, 9000 Rockville Pike, Bethesda, MD 20205.

The Council for International Exchange of Scholars announces the 1985–86 competition for Senior Scholar Fulbright awards for university teaching and postdoctoral research. Awards are offered in all academic fields for periods of 2–10 months in over 100 countries. Prospective applicants may write for applications and additional details on awards, specifying the country and field of interest, to Council for International Exchange of Scholars, 11 Dupont Circle, Suite 300, Washington, D.C. 20036. All applicants must be U.S. citizens, have had college or university teaching experience, and hold the Ph.D. or professional equivalent. Deadlines are June 15, 1984, for American Republics, Australia, and New Zealand, and September 15, 1984, for Africa, Asia, Europe, and the Middle East.

The Indo-U.S. Subcommission on Education and Culture is offering twelve long-term (6–10 months) and nine short-term (2–3 months) awards, without restriction to field, for 1985–86 research in India. Applicants must be U.S. citizens at the postdoctoral or equivalent professional level. Those with limited or no experience in India are especially encouraged to apply. Fellowship terms include $1,500 per month ($350 in dollars and the balance in rupees), an allowance for books and study/travel in India, and international travel for the grantee. In addition, long-term fellows receive international travel for dependents, a dependent allowance of $100–250 per month in rupees, and a supplementary research allowance up to 34,000 rupees. The application deadline is June 15, 1984. Forms and further information an be obtained from the Council for International Exchange of Scholars, Att: Indo-American Fellowships Program, Eleven Dupont Circle, Suite 300, Washington, D.C. 20036. (Telephone: 202+833-4985)

Members of the NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics announce the annual Leonard J. Savage Award of $500 for an outstanding doctoral dissertation in the area of Bayesian Econometrics and Statistics. To be considered for the 1984 Savage Award, a doctoral dissertation must be submitted by the dissertation supervisor before July 1, 1984, accompanied by a short letter summarizing the main results of the dissertation. Dissertations completed after January 1, 1977, are eligible for consideration. The Evaluation Committee will be appointed by the Board of the Leonard J. Savage Memorial Trust Fund (M. H. DeGroot, S. E. Fienberg, S. Geisser, J. B. Kadane, E. E. Leamer, J. W. Pratt, and A. Zellner, chairman). Dissertations and supporting letters should be sent to Professor Arnold Zellner, Graduate School of Business, University of Chicago, 1101. East 58th Street, Chicago, IL 60637.

The winner of the 1983 Award is Paul Garthwaite, "Assessment of Prior Distributions for Normal Linear Models," completed at the University College of Wales, The Evaluation Committee was Bruce M. Hill, chairman, Nicholas, M. Kiefer, David Lane, and Arnold Zellner, *ex officio*.

The first annual award of the Wayne S. Vucinich Prize for the most distinguished monograph in Soviet and East European studies, published in English, was announced in October 1983 by the American Association for the Advancement of Slavic Studies: John R. Lampe, University of Maryland, and Marvin R. Jackson, Arizona State University, coauthors of *Balkan Economic History, 1550–1950; From Imperial Borderlands to Developing Nations* (Indiana University Press). Honorable mention went to Edward J. Brown, *Russian Literature Since the Revolution* (Harvard University Press) and Richard Hellie, *Slavery in Russia 1450–1725* (University of Chicago Press).

The 1982 Irving Fisher Monograph Award was won by Dennis M. Bushe, New York University; Wilhelmus Vijverberg, University of Pittsburgh, received favorable mention. The Frank Taussig Award was not given for 1982; honorable mention was received by Ronald Leaf, University of Minnesota, and by Scott Robinson, University of Wisconsin.

The 1984 International System Dynamics Conference will be held in Oslo, Norway, August 2–5, 1984. The conference seeks to gather a broad international group of professionals and practitioners in the field of system dynamics. For further information, contact Jørgen

Randers, Dean, Norwegian School of Management, Hans Burums vei 30, 1340 Bekkestua, Norway.

The Brookings Institution has received a grant from the National Endowment for the Humanities to partially fund the processing of its Archives. The Archives includes administrative records, research materials relating to the Institution's projects, personal papers of Mr. and Mrs. Robert S. Brookings and Institution officers, and printed material, motion picture films, sound recordings, videotapes, and photographs. The Archives will be closed to researchers for the next eighteen months while the records are being processed. At the completion of the project a finding aid will be published and the holdings made available to researchers. For more information contact Michele F. Pacifico, The Brookings Institution Archives, 1775 Massachusetts Avenue, NW, Washington, D.C. 20036.

The Cliometrics Society (formerly called New Economic History) invites new members. Its tenet is the use of quantitative analyses and economic theory in the study of historical questions. The Society will periodically publish a newsletter, and hold annual meetings. The First World Congress will be held in May 1985. For further information, contact The Cliometrics Society, Department of Economics, Miami University, Oxford, Ohio 45056.

The Institute of North American Studies (Instituto de Estudios Norteamericanos) seeks specialists in the various fields of economics to lecture in Barcelona. The Institute is a binational, nonprofit cultural center founded in 1952 to promote better understanding between Spanish (Catalan in particular) and U.S. culture. A working/lecturing knowledge of Spanish is useful but not essential. Those interested should write Dr. Edward K. Flagler, Director, American Studies Program, Instituto de Estudios Norteamericanos, via Augusta, 123 Barcelona 6, Spain.

The third Symposium on Money, Banking, and Insurance will be held at the University of Karlsruhe, December 12–15, 1984. Papers are invited on monetary theory; interest rates, exchange rates, and inflation; monetary policy; central bank policy; financial theory; financial intermediation; portfolio theory; risk theory and its application to insurance, etc. Submit papers immediately. There is a limited fund for travel grants. Full information can be obtained from Professor Dr. Hermann Göppl, Institut für Entscheidungstheorie und Unternehmensforschung, Universität Karlsruhe (TH), Postfach 6380, D-7500 Karlsruhe 1, West Germany.

The third World Congress for Soviet and East European Studies will be held at the Sheraton Washington Hotel, Washington, D.C., October 30–November 4, 1984; hosted by the American Association for the Advancement of Slavic Studies; and cosponsored by the Association and the International Committee for Soviet and East European Studies. Proposals must be sent immediately to the program Committee Chair, Professor Donald W. Treadgold, School of International Studies, Seattle, WA 98195.

The eighteenth Atlantic Economic Conference will be held in Montreal, October 11–14, 1984. The theme is Bridging the Borders. Those wishing to present a paper should include a submission fee of $25.00 (U.S.), a 500-word summary, and a separate cover listing location of conference plus name of author(s), institution or affiliation, mailing address, telephone number, and number and name of *JEL* catagory for topic. Those wishing to be discussants or chairmen should give the same information. Submit to John M. Virgo, Program Chairman, Atlantic Economic Conference, Southern Illinois University, Box 101, Edwardsville, IL 62026–1001.

The International Association of Energy Economists will hold its sixth annual North American meeting, November 5–7, 1984, at the Fairmont Hotel, San Francisco, California. The program will focus on the outlook for the energy industries through the year 2000, and will include both invited and contributed papers. The emphasis will be on areas of interest to energy industry analysts, especially those concerned with the development of North American Energy markets in general, and U.S. West Coast markets in particular. For more information, contact the Program Chairman, John P. Weyant, Energy Modelling Forum, Terman Engineering, Rm 408, Stanford University, Stanford, CA 94305.

The AEA Committee on the Status of Women in the Economics Profession (CSWEP) session at the 1984 annual meeting of the Southern Economic Association (November 14–16, Atlanta, GA) will be The Impact of Technology Change on the Economic Role of Women. Those wishing to present papers should forward abstracts to Professor Marie Lobue, Department of Economics and Finance, University of New Orleans, New Orleans, LA 70148.

The annual meeting of the Association of Environmental and Resource Economists (AERE) will be held jointly with the AEA in Dallas, Texas, December 28–30, 1984. The AERE will have four contributed papers sessions. Those interested in having papers considered should send two copies of a one-page abstract to V. Kerry Smith, President-Elect, AERE, Department of

Economics, Calhoun Hall, Rm 214, Vanderbilt University, Nashville, TN 37235.

The second annual meeting of the Association of Managerial Economists (AME) will be held in Dallas, Texas, December 28–30, 1984. Three sessions of contributed papers will be featured in conjunction with the ASSA meetings. Theoretical, empirical, and policy studies across a broad spectrum of topics will be included. Both members and nonmembers are invited to submit papers and/or make program suggestions to Professor Mark Hirschey (AME Program Chairman), Graduate School of Business, University of Wisconsin, Madison, WI 53706.

Prospective authors wishing to submit abstracts and papers for the 1985 International Time Series Meetings (date to be confirmed later), should write immediately to O. D. Anderson, 9 Ingham Grove, Lenton Gardens, Nottingham NG7 2LQ, England. Please furnish a self-addressed label.

The *Economics of Education Review* seeks manuscripts related to the entire spectrum of the economics of education, including theoretical, empirical, and policy-oriented papers. Short comments, brief replication studies, and suggestions for book reviews or review essays are welcome. The *EER*, following a one-year hiatus, published Vol. 3, No. 1, in spring 1984. For full information, contract Professor Elchanan Cohn, Editor, *EER*, Department of Economics, College of Business Administration, University, of South Carolina, Columbia, SC 29208.

The journal, *Government and Policy*, invites multidisciplinary and internationally comparative material on social, economic, political, legal, and constitutional analyses of public policy. Submissions should be sent to the Editors: Dr. R. J. Bennett, Department of Geography, University of Cambridge, Downing Place, Cambridge CB2 3EN, or Dr. T. Muller, The Urban Institute, 2100 M Street, NW, Washington, D.C. 20037.

The *Journal of Economic Education* has been reorganized and expanded into a four-section quarterly publication, each section having an associate editor—William Becker: evaluation of teaching techniques, student learning, and related materials; Kalman Goldberg: substantive issues in economics and their incorporation as course content; Karl Case: new teaching techniques or materials; Robin Barlett: economic enrollments, majors, labor markets, and the status of women and minorities. All manuscripts should be sent to Donald W. Paden, Editor, *JEE*, Box 32, David Kinley Hall, 1407 W. Gregory Drive, University of Illinois, Urbana, IL 61801.

Studies in Economic Analysis is a biannual, student-edited journal currently expanding its scope and soliciting research articles from both established economists and students. There is no submission fee. Submit manuscripts to or request complete format and style requirements from the Editors, *SEA*, Department of Economics, College of Business Administration, University of South Carolina, Columbia, SC 29208.

Trinity College seeks books—donations of quality hardcover and paper—in all areas of scholarly interest, but especially in the social sciences. Gifts are tax deductible; donors will receive individual letters of acknowledgement. Send gifts and inquiries to The Library Director, Trinity College, Burlington, VT 05401.

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air far between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for application to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 800 Third Avenue, New York, NY 10022, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

Deaths

Otto Eckstein, Chairman, Data Resources, Inc., Lexington, MA, March 22, 1984.

George W. Jennings, professor emeritus, Virginia Commonwealth University, January 21, 1984.

Donald D. Kennedy, diplomat and educator, Portland, Oregon, June 14, 1983.

Retirements

George L. Bach, Stanford Business School, September 1, 1983.

Samuel C. Kelley, professor of economics, Ohio State University, February 1, 1984.

Karl F. Treckel, professor of economics, Kent State University, May 19, 1984.

Foreign Scholars

Athena P. Kottis, Athens School of Economics and Business: visiting scholar, George Washington University, 1983–84.

George C. Kottis, Athens School of Economics and Business: adjunct (research) professor, American University, 1983–84.

Promotions

James M. Boughton: assistant chief, external adjustment division, research department, International Monetary Fund, October 1, 1983.

Hui-shyong Chang: professor of economics, University of Tennessee-Knoxville, September 1, 1983.

Don P. Clark: associate professor of economics, University of Tennessee-Knoxville, September 1, 1983.

William W. Damon: professor of economics and business administration, Vanderbilt University, September 1983.

Henry W. Herzog, Jr.: professor of economics, University of Tennessee-Knoxville, September 1, 1983.

J. Huston McCulloch: professor of economics, Ohio State University, July 1, 1983.

Howard P. Marvel: professor of economics, Ohio State University, July 1, 1983.

Takahiro Miyao: professor of economics, University of Southern California, September 1, 1983.

Nina Ramondelli: vice president and economist, Chase Manhattan Bank, November, 1, 1983.

Vaman Rao: professor of economics, Western Illinois University, August 1982.

A. H. Studenmund: professor of economics and finance, Occidental College, July 1983.

Roy H. Webb: research officer, Federal Reserve Bank of Richmond, January 1, 1983.

Administrative Appointments

M. Arshad Chawdhry: chairman, business and economics department, California University of Pennsylvania, June 1983–May 1985.

Jeffrey I. Chapman: director, Sacramento Public Affairs Center, University of Southern California, July 1, 1983.

Gary A. Latanich: assistant dean, College of Business, Arkansas State University, January 1, 1984.

Frank W. Puffer: chairman, department of economics, Clark University, January 1, 1984.

Jorge Salazar-Carrillo: chairman, department of economics, Florida International University, August 1983–August 1986.

Appointments

Robin L. Allen, Northwestern University: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, October 1983.

Kenneth C. Baseman, Owen, Greenhalgh and Myslinski, Economists, Inc.: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, May 1983.

Michael G. Baumann, Harvard University: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, September 1983.

Richard Cantor, Johns Hopkins University: assistant professor of economics, Ohio State University, October 1, 1983.

Richard N. Clarke, University of Wisconsin-Madison: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, June 1983.

Robert Driskill, University of California-Davis: associate professor of economics, Ohio State University, October 1, 1983.

Elaine M. Gilby, University of Wisconsin: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, September 1983.

Paul E. Godek, University of Chicago: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, August 1983.

Steven L. Green: instructor in economics, Vanderbilt University, September 1983.

Gikas A. Hardouvelis: assistant professor of economics, Barnard College, July 1, 1983.

Richard J. Kent, Jr.: associate professor of economics, Kent State University, fall 1983.

Niccie L. McKay, Massachusetts Institute of Technology: assistant professor of economics, Texas A&M University, September 1983.

Nelson Mark, University of Chicago: assistant professor of economics, Ohio State University, October 1, 1983.

Patricia Reagan, University of Rochester: assistant professor of economics, Ohio State University, October 1, 1983.

Michael Tretheway: assistant professor of transportation, Faculty of Commerce and Business Administration, University of British Columbia, July 1, 1983.

Donald R. Williams: assistant professor of economics, Kent State University, fall 1983.

Kenneth Wolpin, Yale University: professor of economics, Ohio State University, October 1, 1983.

Leaves for Special Appointments

Russell G. Pounds, Iowa State University: Zambia, January 4, 1984–July 7, 1984.

Gian S. Sahota, Vanderbilt University: chief technical advisor, United Nations Development Program, Republic of Panama, July 1983.

Frederick R. Warren-Boulton, Washington University: Director, Economic Policy Office, Antitrust Division, U.S. Department of Justice, September 1983.

Resignations

Stanley W. Black, Vanderbilt University: University of North Carolina-Chapel Hill, September 1983.

Trent E. Boggess, Kent State University, August 19, 1983.

Vladimir J. Simunek, Kent State University, December 20, 1983.

Neil A. Stanley, Iowa State University, December 31, 1983.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment: her new title (if any), new institution and the date at which the change will occur.

C. Type each item on a separate 3×5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

All items and information should be sent to the Assistant Production Editor, *American Economic Review*, Room 8279, Bunche Hall, University of California, Los Angeles, CA 90024.

NOTICE TO ALL GRADUATE DEPARTMENTS

The December 1984 issue of the *Review* will carry the eighty-first list of doctoral dissertations in political economy in American universities and colleges. The list will give recipients and titles of doctoral degrees conferred during the academic year terminating June 1984. This announcement is an invitation to send us information for the preparation of the list.

By June 30, please send us this information on 3×5 cards, conforming to the style shown below, one card for each individual. Please indicate by a classification number in the right-hand corner the field in which the thesis should be classified. The classification system is that used by the *Journal of Economic Literature* and printed in every issue.

All items and information should be sent to the Assistant Production Editor, *American Economic Review*, Room 8279, Bunche Hall, University of California, Los Angeles, CA 90024.

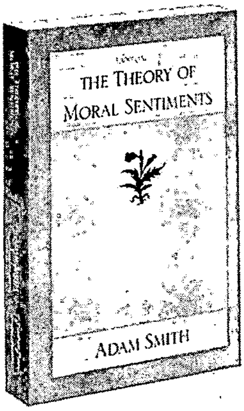
***JEL* Classification No.** ________

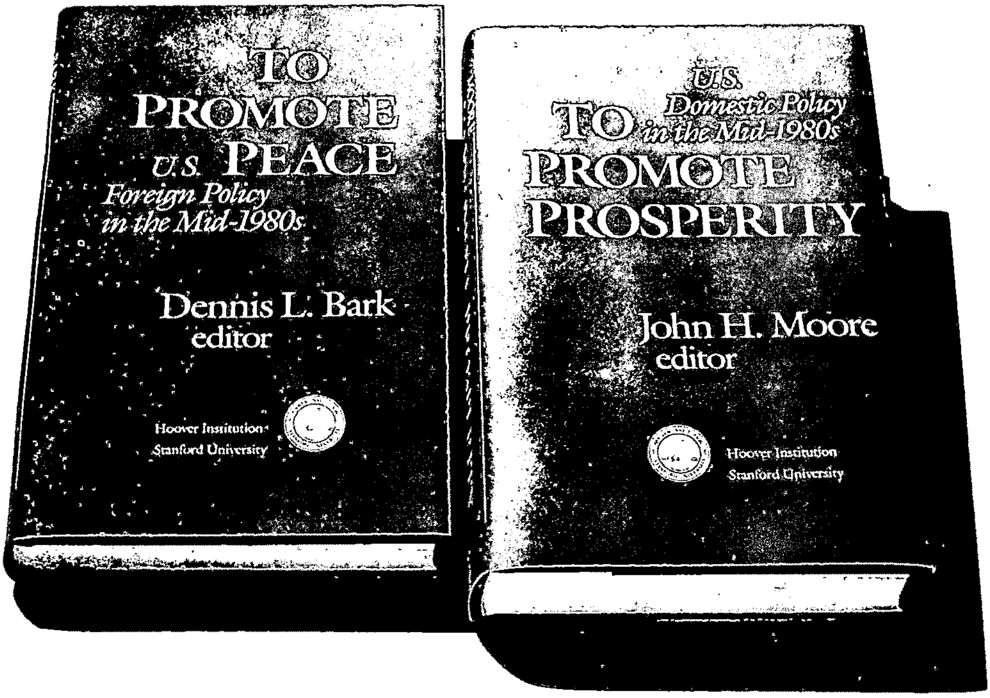
Name: LAST NAME IN CAPS: First Name, Initial

Institution Granting Degree: ________________________

Degree Conferred (Ph.D. or D.B.A.) ____________ **Year** ____________

Dissertation Title: ________________________


THE THEORY OF
MORAL SENTIMENTS
ADAM SMITH


TO PROMOTE U.S. PEACE
Foreign Policy in the Mid-1980s
Dennis L. Bark editor
Hoover Institution Stanford University
U.S. Domestic Policy in the Mid-1980s
TO PROMOTE PROSPERITY
John H. Moore editor
Hoover Institution Stanford University

AMERICAN ECONOMIC ASSOCIATION

1984 ANNUAL MEMBERSHIP RATES

Membership includes:

—a subscription to both *The American Economic Review* (quarterly) plus *Papers and Proceedings* and the *Journal of Economic Literature* (quarterly).

- Regular member with rank of assistant professor or lower or annual incomes of $15,840 or less $33.00
- Regular member with rank of associate professor or annual incomes of $15,840/$26,400 .... $39.60
- Regular member with rank of full professor or annual income above $26,400 .... $46.20
- Junior member (available to registered students for three years only). Student status must be certified by your major professor or school registrar $16.50
- In Countries other than the U.S.A., Add $9.20 to cover postage.
- Family member (second membership without publications; two or more living at same address) ..... $ 6.60

Please begin my issues with:

□ **March** □ **June** (*Includes Papers and Proceedings*) □ **September** □ **December**

_______________ _______________ _______________
First Name and Initial Last Name Suffix

____
Address Line 1

____
Address Line 2

____
City

_______________ _______________
State or Country Zip/Postal Code

MAJOR FIELDS (TWO ONLY)
LIST FIELDS WITH WHICH YOU CURRENTLY IDENTIFY. SELECT FIELD CODE FROM *JEL*, "Classification System for Books."

□ □

PLEASE TYPE OR PRINT INFORMATION ABOVE; PLEASE SEND CHECK OR MONEY ORDER PAYABLE IN U.S. DOLLARS. CANADIAN AND FOREIGN PAYMENTS MUST BE IN THE FORM OF A U.S. DOLLAR DRAFT ON A NEW YORK BANK.

Endorsed by (AEA member) _______________

Below for Junior Members Only

I certify that the person named above is enrolled as a student at _______________

____

____
Authorized Signature

PLEASE SEND WITH PAYMENT TO:

AMERICAN ECONOMIC ASSOCIATION
1313 21ST AVENUE SOUTH, SUITE 809
NASHVILLE, TENNESSEE 37212-2786
U.S.A.